Journal of Multivariate Analysis 101 (2010) 765–788

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Nonparametric tests for conditional independence in two-way contingency tables

Gery Geenens [a,b,*], Léopold Simar [a]

[a] Institut de Statistique, Université catholique de Louvain, Belgium
[b] Department of Mathematics and Statistics, University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 2 January 2008
Available online 7 January 2010

AMS subject classifications:
primary 62H17
62G08
secondary 62H20
62G10

Keywords:
Two-way contingency tables
Chi-square test
Likelihood ratio test
Nonparametric regression
Conditional independence

ABSTRACT

Testing for the independence between two categorical variables R and S forming a contingency table is a well-known problem: the classical chi-square and likelihood ratio tests are used. Suppose now that for each individual a set of p characteristics is also observed. Those explanatory variables, likely to be associated with R and S, can play a major role in their possible association, and it can therefore be interesting to test the independence between R and S conditionally on them. In this paper, we propose two nonparametric tests which generalise the chi-square and the likelihood ratio ideas to this case. The procedure is based on a kernel estimator of the conditional probabilities. The asymptotic law of the proposed test statistics under the conditional independence hypothesis is derived; the finite sample behaviour of the procedure is analysed through some Monte Carlo experiments and the approach is illustrated with a real data example.
© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Let R and S be two categorical variables, with r and s levels respectively, and consider a sample of n individuals for which R and S are known. A contingency table is built by cross-classifying the sample with respect to the levels of R and S. The quantities of interest facing such a table are typically the joint probability distribution $\pi = \{\pi_{ij} : 1 \leq i \leq r, 1 \leq j \leq s\}$ of R and S, with

$$\pi_{ij} = \mathbb{P}(R = i, S = j)$$

the probability that a given individual belongs to the cell (i, j) of the table, and the ensuing marginal probabilities $\pi_{i\cdot} = \mathbb{P}(R = i)$ and $\pi_{\cdot j} = \mathbb{P}(S = j)$. All those quantities are easily estimated from the sample proportions $\hat{p}_{ij}$. A fundamental question in this context is whether R and S are independent or not, which is formalised as

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall (i, j). \tag{1.1}$$

Most of the testing procedures for this hypothesis rely on a divergence criterion between $\{\hat{p}_{ij}\}$ and $\{\hat{p}_{i\cdot}\hat{p}_{\cdot j}\}$, such as the well-known Pearson's chi-square statistic and the likelihood ratio test statistic. Under the null hypothesis, these two statistics are asymptotically equivalent and follow a χ^2 distribution with $(r-1)(s-1)$ degrees of freedom, which allows us to define an asymptotic rejection criterion for H_0.

* Corresponding address: Institut de Statistique, Université catholique de Louvain, voie du Roman Pays, 20, 1348 Louvain-la-Neuve, Belgium.
E-mail addresses: ggeenens@unimelb.edu.au (G. Geenens), leopold.simar@uclouvain.be (L. Simar).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2009.12.017

One limitation of those classical procedures is that the distribution π is assumed to be the same for each individual. Yet, in most of the situations, each individual possesses some characteristics of his own, say a vector X of explanatory variables, which ought to influence R, S, or both, and could consequently play a major role in their possible association. The main idea of this paper is therefore to extend the classical ideas in order to be able to take into account a possible heterogeneity in the population of interest, by clearing the possible effect of X from the analysis, and by doing so to go further in the study of the association observed in the concerned contingency table. For example, imagine that a classical test of independence (χ^2 or likelihood ratio) emphasises a significant association between R and S. Then, one could wonder if this association is not "artificially" implied by some external factors, which would be strongly related to both variables. On the contrary, if the classical tests fail to stress any significant association on a global scale, one could look for some 'local' association, by comparing like subjects to like subjects. Working conditionally to the considered external factors would probably give answers to those interrogations, as it is well known in the statistical theory that working conditionally on a random variable is very much like removing the effect that this variable can have on the analysis which is being carried out, seeing it as fixed to some value. It must be made clear that a test procedure for conditional independence is not really a concurrent or a superior version of the usual unconditional independence tests, but rather a complementary analysis which could shed some new light on the observed phenomenon. See Section 4 where conditional and unconditional independence are compared and related to each other. Now, it seems natural to base our procedure on the joint distribution of R and S conditional on X, i.e. $\pi(x) = \{\pi_{ij}(x) : 1 \leq i \leq r,\ 1 \leq j \leq s\}$, with

$$\pi_{ij}(x) = \mathbb{P}\left(R = i, S = j | X = x\right),$$

aiming at testing for the conditional independence of those two categorical variables given the set of extra covariates. If $S_X \subset \mathbb{R}^p$ is the support of X, we naturally define this conditional independence hypothesis as

$$H_0 : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \quad \forall x \in S_X, \forall (i, j). \tag{1.2}$$

It is here probably worth noting that, in the contingency tables theory (see a.o. [1,2]), the term "conditional independence" often refers to the independence of R and S given T, where R, S and T are three categorical variables on an equal footing with each other and forming a three-way contingency table. Our idea is different, as we make a clear distinction between the variables R and S whose association is of interest and the explanatory vector X. However, if X is a discrete random vector, the same type of methodology can apply (see e.g. the Cochran–Mantel–Haenszel test [3] or [1, Section 4.3.4]). Therefore, throughout this paper, we only focus on the case where X is continuous. To our knowledge, this work is the first to address the problem of directly testing for (1.2) in this case.

Now, such a procedure has the obvious need of reliable estimates of $\pi(x)$, from a sample of individuals drawn at random from the population of interest. This paper concentrates on a nonparametric estimation of this vector of functions. The motivation to favour this kind of methods rather than parametric ones is exposed in Section 2, as well as theoretical results about the proposed nonparametric estimator of $\pi(x)$. In Section 3, the test procedures are described, and the asymptotic distribution of the test statistics are derived. Section 4 highlights some interesting observations about how hypotheses (1.1) and (1.2) are related. Last sections provide a simulation study in Section 5, a real data example in Section 6, and the concluding remarks in Section 7.

2. Nonparametric estimation of $\pi(x)$

Define the the random vector

$$Z = (Z^{(11)}, Z^{(12)}, \ldots, Z^{(rs)})^t,$$

with $Z^{(ij)}$ taking the value 1 if the individual belongs to cell (i, j) and 0 otherwise. The components of Z are indexed by the pairs (ij), such that the index (ij) denotes the $((i-1)s+j)$th component of the vector, for convenience. This will also be the case for most of the vectors defined in the sequel. In the same spirit, $\sum_{ij=11}^{rs}$, or simply $\sum_{ij}$, will often be written in place of $\sum_{i=1}^{r}\sum_{j=1}^{s}$. The following assumption formalises the context that will be considered throughout the paper:

Assumption 2.1. The sample is described by $\{(X_k, Z_k)\}_{k=1}^n$, which forms a sequence of i.i.d. replications of $(X, Z) \in S_X \times \{z \in \{0, 1\}^{rs} : \sum_{ij=11}^{rs} z^{(ij)} = 1\}$, a random vector such that $Z|X$ follows a multinomial distribution with parameters $(1, \pi(X))$.

Clearly, this assumption extends in a straightforward way the classically assumed multinomial sampling to the case where some external factors are likely to influence R and S. It directly follows from Assumption 2.1 that

$$\pi_{ij}(x) = \mathbb{E}(Z^{(ij)}|X = x), \tag{2.1}$$

so that the estimation of $\pi(x)$ is clearly nothing else but a multiresponse regression problem.

2.1. Motivation for a nonparametric approach

Generalised Linear Models have therefore been proposed to estimate $\pi(x)$. The most popular is probably the multivariate logistic regression model, introduced by McCullagh and Nelder [4, Section 6.5.4] and discussed in [5]. This model can be written in the general form

$$C^t \log(L\pi(X)) = \Theta X \tag{2.2}$$

where L and C are appropriately chosen matrix of 0, 1 and -1 only, and Θ a matrix of unknown parameters. As illustration, in the case $r = s = 2$, this yields

$$\text{logit}(\pi_{1\cdot}(x)) = \theta_R^t x; \qquad \text{logit}(\pi_{\cdot 1}(x)) = \theta_S^t x; \qquad \log\left(\frac{\pi_{11}(x)\pi_{22}(x)}{\pi_{12}(x)\pi_{21}(x)}\right) = \theta_{RS}^t x. \tag{2.3}$$

Other similar parametric models could be mentioned, e.g. log-linear regression models or multivariate probit models, and their generalisations. See a.o. [6–8] or [9]. These models take advantage of the properties of the parametric procedures. In particular, testing for (1.2) amounts to some simple parametric inference, as testing for $\theta_{RS} = 0$ in model (2.3). However, some important issues can be pointed out. First, those models suffer from their usual lack of flexibility. For example, they do not allow non-monotonic links between the linear predictor and the corresponding conditional probabilities. Second, the binary character of the responses $Z^{(ij)}$ leads to difficulties in analysing scatter-plots. As observing the shape of the cloud of points is often the primary tool for defining a reliable parametric pattern for a regression function, the risk of misspecification is much more important here than in a classical regression context, as already emphasised in [10]. In addition, as the assumed conditional probabilities have to be linearly related by $\sum_{ij}\pi_{ij}(x) \equiv 1$, not one, but rs scatter-plots have to be simultaneously analysed, and the exercise is still riskier. Third, again due to the 0–1 responses, standard residuals-based model checking techniques are not adapted: residuals are here simply the complements to 0 or 1 of the link functions assumed by the model, so that their representation is not really informative. Finally, a fourth point is that in model (2.2) for example, Θ is a matrix of size $(rs-1)\times(p+1)$. The number of parameters to be estimated is thus $(rs-1)(p+1)$, possibly already very large for a moderate size of the table and a moderate number of covariates. In this case, the Maximum Likelihood estimation of these parameters relies on a high-dimensional optimisation problem, from which practical difficulties frequently arise.

To get around those drawbacks, we propose to nonparametrically estimate the conditional probabilities $\pi_{ij}(x)$. Besides, the development which follows shows that the proposed method is particularly well adapted to the considered setting. The use of nonparametric regression techniques for binary data was first studied by Copas [11]. Later, [12–15] a.o., used a Nadaraya–Watson estimator (NW) in this context. Note that the Local Linear estimator (LL), known to theoretically outperform the NW estimator (see a.o. [16]), does not seem to be suitable here. Indeed, the LL estimator need not belong to the range of the observed responses, contrary to the NW estimator. This would be seriously problematic in our framework as the estimated probabilities could be found negative or greater than 1. We therefore focus on a Nadaraya–Watson-like estimator.

2.2. Assumptions and definition of the estimator

For seek of brevity, we here only address the case where X is a scalar continuous variable. We refer to [17] for results and comments in the multivariate case. Let K be a kernel function and h a bandwidth, the usual parameters in nonparametric regression. The following regularity conditions are assumed to hold.

Assumption 2.2. The support S_X of X is compact, and X admits a density f such that for any $x \in S_X$, $0 < f(x) < \infty$ and $f(x)$ has three bounded derivatives;

Assumption 2.3. For any $x \in S_X$, any $\pi_{ij}(x)$ is bounded away from 0 and from 1 and has three bounded derivatives;

Assumption 2.4. The kernel function K is a bounded symmetric Lipschitz continuous probability density on $[-1, 1]$;

Assumption 2.5. The bandwidth sequence $h \doteq h_n$ is such that $h \to 0$ and $nh \to \infty$.

These assumptions are standard in nonparametric regression, except the required three derivatives of the concerned functions. This extra smoothness actually makes some results hold uniformly in x, as it will be required in Section 3. Note also that Assumption 2.2 covers in a sense the case where X is non-random (fixed design), as it is well known in the kernel regression theory that this situation amounts to consider a uniform design, that is to take f constant (obviously depending on S_X) in the results. See [18, Sections 5.3.1–5.3.2].

Then, from (2.1), the Nadaraya–Watson estimator of $\pi_{ij}(x)$ is given by

$$\hat{p}_{ij}(x) = \sum_{k=1}^{n} W_h(x, X_k) Z_k^{(ij)}, \tag{2.4}$$

with

$$W_h(x, X_k) = \frac{K\left(\frac{x-X_k}{h}\right)}{\sum_{k'=1}^{n} K\left(\frac{x-X_{k'}}{h}\right)}.$$

See in addition that Assumption 2.1 also implies that

$$\mathbb{V}\mathrm{ar}(Z^{(ij)}|X = x) = \pi_{ij}(x)(1 - \pi_{ij}(x)).$$

Therefore, Assumptions 2.1–2.5 are sufficient to ensure that the usual theoretical properties of the NW estimator directly apply to estimator (2.4).

2.3. A common bandwidth

In particular, the asymptotically optimal value of h to estimate $\pi_{ij}(x)$ is known to be

$$h_{\mathrm{opt}}^{(ij)} = \left(\frac{\nu_0 \int_{S_X} \pi_{ij}(x)(1 - \pi_{ij}(x))\mathrm{d}x}{\mu_2^2 \int b_{ij}^2(x) f(x)\mathrm{d}x}\right)^{1/5} n^{-1/5},$$

with $\mu_q = \int x^q K(x)\mathrm{d}x$, $\nu_q = \int x^q K^2(x)\mathrm{d}x$ and $b_{ij}(x)$ defined in (2.9) below. This result, although not applicable in practice since based on the unknown $\pi_{ij}(x)$ and $f(x)$, seems to indicate anyway that it might be preferable to use a different bandwidth $h^{(ij)}$ for each $\hat{p}_{ij}(x)$.

Nevertheless, this would cause some undesirable features in further developments. Firstly, the $\{\hat{p}_{ij}(x)\}$ would not necessarily sum to one:

$$\sum_{ij} \hat{p}_{ij}(x) = \sum_{ij}\sum_{k} W_{h^{(ij)}}(x, X_k) Z_k^{(ij)} \not\equiv 1,$$

as it would be the case if a common bandwidth h was used. Indeed, in this latter case, we would have

$$\sum_{ij} \hat{p}_{ij}(x) = \sum_{ij}\sum_{k} W_h(x, X_k) Z_k^{(ij)} = \sum_{k} W_h(x, X_k) \sum_{ij} Z_k^{(ij)} \equiv 1,$$

since $\sum_k W_h(x, X_k) \equiv 1$ and one and only one of the $Z_k^{(ij)}$ is 1. Secondly, the marginal estimate of $\pi_{i\cdot}(x)$, based on observations $\{Z_k^{(i\cdot)}\}$ and on a bandwidth $h^{(i)}$, would not be equal to the sum of the estimates $\hat{p}_{ij}(x)$:

$$\hat{p}_i(x) \doteq \sum_{k} W_{h^{(i)}}(x, X_k) Z_k^{(i\cdot)} \neq \sum_{j}\sum_{k} W_{h^{(ij)}}(x, X_k) Z_k^{(ij)} = \hat{p}_{i\cdot}(x).$$

Again, this would be the case if the same bandwidth was used in each cell. Finally, define the vector

$$\hat{p}(x) = \left(\hat{p}_{11}(x), \hat{p}_{12}(x), \ldots, \hat{p}_{r(s-1)}(x), \hat{p}_{rs}(x)\right)^t, \tag{2.5}$$

and see that the use of a common bandwidth allows this whole vector to be computed in one time, in a very fast and easy way. Indeed, with

$$\mathcal{W}_h(x) = (W_h(x, X_1), \ldots, W_h(x, X_n))^t$$

and

$$\mathcal{Z} = \begin{pmatrix} Z_1^{(11)} & Z_1^{(12)} & \cdots & Z_1^{(rs)} \\ Z_2^{(11)} & Z_2^{(12)} & & \\ \vdots & & \ddots & \vdots \\ Z_n^{(11)} & & \cdots & Z_n^{(rs)} \end{pmatrix},$$

we have directly that

$$\hat{p}(x) = \mathcal{Z}^t \mathcal{W}_h(x). \tag{2.6}$$

The rs conditional probabilities do not need to be estimated by rs different nonparametric estimators, but are readily given by a simple matrix product. This fact can be of importance in practice, for example with regard to the computing time.

Therefore, in order that the $\{\hat{p}_{ij}(x)\}$ keep the essential properties of the underlying $\{\pi_{ij}(x)\}$, as well as for practical facility, it seems judicious to use an appropriately chosen common bandwidth h in every cell. We propose to define the theoretical

optimal common bandwidth h_{opt} as the value which minimises the sum of the asymptotic Integrated Mean Squared Error of each $\hat{p}_{ij}(x)$. Straightforward calculations show that

$$h_{\text{opt}} = \left(\frac{\nu_0 \int_{S_X} (1 - \sum_{ij} \pi_{ij}^2(x)) \mathrm{d}x}{\mu_2^2 \int \sum_{ij} b_{ij}^2(x) f(x) \mathrm{d}x} \right)^{1/5} n^{-1/5}, \tag{2.7}$$

theoretical expression from which practical bandwidth choice rules could be derived, similarly to usual ones (plug-in methods, cross-validation, ...).

To sum up, the proposed nonparametric estimator (2.6) of the conditional distribution of interest is very fast and easy to compute, and fulfils any constraint imposed by the context without extra work: $\hat{p}_{ij}(x) \in [0, 1]$ and $\sum_{ij} \hat{p}_{ij}(x) \equiv 1$, automatically. It therefore adapts especially well to the considered framework, without the need for any structural prior assumption on the design.

2.4. Asymptotic properties

Define the "interior" of the support S_X as $S_X^{(h)} \doteq \{x \in S_X : m_X + h \leq x \leq M_X - h\}$, where m_X and M_X are the lower and the upper bound of S_X. Such a set needs to be defined as it is well known that the behaviour of the Nadaraya–Watson estimator differs when computed at points close to the boundary of the support. In the sequel, the observations which do not belong to this interior set will be trimmed, although a more elaborate version of the NW estimator, designed by usual way to automatically remedy this problem (see e.g. [19]), could be used to avoid this trimming. Then, from the classical theory of kernel regression, if $h = O(n^{-1/5})$, we have, for any fixed $x \in S_X^{(h)}$ and $\forall (i, j)$:

$$(nh)^{1/2} \left(\hat{p}_{ij}(x) - \pi_{ij}(x) - h^2 b_{ij}(x) \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{ij}^2(x)), \tag{2.8}$$

where

$$b_{ij}(x) = \frac{1}{2} \kappa_2 \left(\pi_{ij}''(x) + \pi_{ij}'(x) \frac{f'(x)}{f(x)} \right) \quad \text{and} \quad \sigma_{ij}^2(x) = \nu_0 \frac{\pi_{ij}(x)(1 - \pi_{ij}(x))}{f(x)}. \tag{2.9}$$

In addition, due to the assumed multinomial sampling scheme (Assumption 2.1), it can easily be shown that the asymptotic covariance between $\hat{p}_{i_1 j_1}(x)$ and $\hat{p}_{i_2 j_2}(x)$ equals $-\nu_0 \frac{\pi_{i_1 j_1}(x) \pi_{i_2 j_2}(x)}{nhf(x)}$, for $(i_1, j_1) \neq (i_2, j_2)$. Therefore, defining vectors

$$\begin{aligned} \pi(x) &= \left(\pi_{11}(x), \pi_{12}(x), \ldots, \pi_{r(s-1)}(x), \pi_{rs}(x) \right)^t \\ b(x) &= \left(b_{11}(x), b_{12}(x), \ldots, b_{r(s-1)}(x), b_{rs}(x) \right)^t, \end{aligned} \tag{2.10}$$

the vector analogue of (2.8) can be shown to be

$$(nh)^{1/2} \left(\hat{p}(x) - \pi(x) - h^2 b(x) \right) \xrightarrow{\mathcal{L}} \mathcal{N} \left(0, \frac{\nu_0}{f(x)} \left(\mathrm{diag}(\pi(x)) - \pi(x) \pi(x)^t \right) \right) \tag{2.11}$$

with $\mathrm{diag}(\pi(x))$ being the diagonal matrix built on the elements of $\pi(x)$.

Finally, the following result will also be useful in the sequel. Denote the convolution of the kernel K with itself by $\nu_0(u) \doteq (K * K)(u)$.

Lemma 2.1. *Under Assumptions 2.1–2.5, we have, for any x_1 and $x_2 \in S_X^{(h)}$,*

$$\mathbb{C}\mathrm{ov} \left(\hat{p}_{ij}(x_1), \hat{p}_{i'j'}(x_2) \right) = \frac{\nu_0(\delta)}{nhf(x_1)} \pi_{ij}(x_1) \left(\delta_{i,i'} \delta_{j,j'} - \pi_{i'j'}(x_1) \right) (1 + O(h)), \tag{2.12}$$

as $n \to \infty$, with $\delta = \frac{x_1 - x_2}{h}$ and $\delta_{i,i'}$ being the Kronecker delta. Besides, the order of the remainder term holds uniformly in $x_1, x_2 \in S_X^{(h)}$.

Proof. See Appendix. □

Remark 2.1. As $\nu_0(\delta) = 0$ once $|\delta| > 2$ by Assumption 2.4, the covariance between $\hat{p}_{ij}(x_1)$ and $\hat{p}_{i'j'}(x_2)$ is zero once $|x_1 - x_2| > 2h$, which is obvious as the estimations at x_1 and at x_2 are then based on disjoint sets of independent observations. On the other hand, if $|x_1 - x_2| < 2h$, $\nu_0(\delta) > 0$ quantifies the lapping of the two kernels centred at x_1 and x_2, that is in a sense the weight of observations influencing at the same time $\hat{p}_{ij}(x_1)$ and $\hat{p}_{i'j'}(x_2)$. This coefficient can thus be interpreted as the amount of information shared by $\hat{p}_{ij}(x_1)$ and $\hat{p}_{i'j'}(x_2)$, and therefore plays the central role in the expression of their covariance.

Remark 2.2. The covariance expression (2.12) does not seem to be symmetric in x_1 and x_2, which could appear a bit surprising. However, as it is not identically zero only if $|x_1 - x_2| < 2h$ as highlighted by the previous remark, that is if $x_1 \to x_2$ since $h \to 0$, the continuity of f and $\{\pi_{ij}\}$ allows us to validate it: either $\nu_0(\delta)$ is zero, or $\pi_{ij}(x_2) = \pi_{ij}(x_1) + O(h)$, and a "symmetrised" version of (2.12) could easily be written without changing the result.

3. Testing for conditional independence

3.1. Test statistics

As introduced in Section 1, the overall idea of this paper is to generalise the chi-square and the likelihood ratio criteria to cases where a vector of covariates is involved. Ideally, for any x in S_X, a pointwise divergence criterion between the estimated joint conditional distribution of R and S given $X = x$ and the product of the marginal conditional distributions of R and S given $X = x$ would be computed, and then integrated with respect to x, in order to evaluate this divergence on the whole support S_X. For instance, in view of (2.8) and (2.9), a natural generalisation of the Pearson's χ^2-criterion is

$$\int_{S_X} \frac{nhf(x)}{\nu_0} \sum_{ij=11}^{rs} \frac{\left(\hat{p}_{ij}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)} f(x)\mathrm{d}x. \tag{3.1}$$

Define the vector $v(x) = (v_{11}(x), v_{12}(x), \ldots, v_{r(s-1)}(x), v_{rs}(x))^t$, with

$$v_{ij}(x) = \frac{\sqrt{nhf(x)}\left(\hat{p}_{ij}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)}{\sqrt{\nu_0}\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}}, \tag{3.2}$$

such that (3.1) equals $\int_{S_X} \|v(x)\|^2 f(x)\mathrm{d}x$. This vector depends on the unknown $\pi(x)$ and $f(x)$, which have to be estimated. The vector $\pi(x)$ is naturally estimated by $\hat{p}(x)$, while $f(x)$ can be estimated by the usual nonparametric kernel density estimator[1]

$$\hat{f}(x) = \frac{1}{nh}\sum_{k=1}^{n} K((x - X_k)/h), \tag{3.3}$$

see a.o. [18]. Write $\hat{v}(x)$ for the estimated version of $v(x)$, that is the vector whose components are

$$\hat{v}_{ij}(x) = \frac{\sqrt{nh\hat{f}(x)}\left(\hat{p}_{ij}(x) - \hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)\right)}{\sqrt{\nu_0}\sqrt{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}}.$$

Remark 3.1. When $\hat{p}_{i\cdot}(x) = 0$ or $\hat{p}_{\cdot j}(x) = 0$, the undetermined $\hat{v}_{ij}(x)$ is set to 0. This can be intuitively justified the following way: the fact that an individual characterised by this x has very few chances to fall into ith category of R, or into jth category of S, does not bring evidence against hypothesis H_0, so that no contribution to the divergence criterion has to be considered. This can be seen as a natural trimming of the observations.

Finally, observing that $\int_{S_X} \|v(x)\|^2 f(x)\mathrm{d}x = \mathbb{E}(\|v(X)\|^2)$, a natural estimation of the integral is

$$V^2 = \frac{1}{n}\sum_{k=1}^{n} \|\hat{v}(X_k)\|^2 \mathbb{1}(X_k \in S_X^{(h)}), \tag{3.4}$$

hopefully close to $\mathbb{E}(\|v(X)\|^2)$ by the Law of Large Numbers, and given that $\hat{v}$ is a consistent estimate of v. The trimming is added in order to use results such as (2.11) to derive the asymptotic distribution of any $\|\hat{v}(X_k)\|^2$ entering the sum.

Similarly, we propose to generalise the Likelihood Ratio test statistic by

$$\hat{G}(x) = 2\frac{nh\hat{f}(x)}{\nu_0}\sum_{i,j} \hat{p}_{ij}(x)\log\frac{\hat{p}_{ij}(x)}{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)} \tag{3.5}$$

for any fixed x, and to gather results for all x by taking the empirical mean

$$G = \frac{1}{n}\sum_{k=1}^{n} \hat{G}(X_k)\mathbb{1}(X_k \in S_X^{(h)}).$$

Note that in (3.5), the undetermined $\hat{p}_{ij}(x)\log\frac{\hat{p}_{ij}(x)}{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}$ is set to zero if $\hat{p}_{ij}(x) = 0$, for the same reasons as those explained in Remark 3.1, and given that $\lim_{y\to 0} y\log y = 0$.

[1] For simplicity, the same bandwidth h is used here for the estimation of $\hat{p}_{ij}(x)$ and $\hat{f}$, but this is not required. In fact, any bandwidth g such that $h = O(g)$ could be used for this estimation, check the proof of Lemma 3.2.

3.2. Bias treatment

An undesirable feature observed in the asymptotic behaviour of $\hat{p}(x)$ is the presence of the bias term. Indeed, see from (2.8) that if $h \sim n^{-1/5}$ as suggested by (2.7), the asymptotic distribution of $\sqrt{nh}(\hat{p}_{ij}(x) - \pi_{ij}(x))$ is not centred at zero. This fact is seriously problematic in the context of the considered tests. Look for example at (3.2): even if H_0 holds, we have

$$\mathbb{E}(v_{ij}(x)) \not\equiv 0,$$

and the divergence criterion fails to behave as it should. It is therefore needed to correct it for the bias, which is common in procedures where nonparametric regression is involved, and usually carried out via two methods. The first is to estimate the bias, and then to proceed to an explicit bias correction where it is needed (see a.o. [20,21] or [22] for this way-of-doing in other situations). The second method is to proceed via undersmoothing. As expressed by (2.7), the optimal bandwidth is $h \sim n^{-1/5}$. However, if it is taken $o(n^{-1/5})$, the bias term asymptotically vanishes in (2.11). The idea is thus to voluntarily work with a bandwidth which is not optimal, in order to implicitly deal with the bias, of course at the expense of a more important variance. We then get

$$(nh)^{1/2}\left(\hat{p}(x) - \pi(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\nu_0}{f(x)}\left(\mathrm{diag}(\pi(x)) - \pi(x)\pi(x)^t\right)\right) \quad \forall x \in S_X^{(h)}, \tag{3.6}$$

once $h = o(n^{-1/5})$. This implicit correction will be favoured in this paper. Indeed, explicit estimation of the bias requires extra work, all the more since in our framework, not one but *rs* bias terms should be estimated. Also, in various situations, undersmoothing is often seen to yield better results than explicit correction, see [23] or [24]. And last, if a correction term was added to the primary estimates, we should lose the important properties of the $\{\hat{p}_{ij}(x)\}$ described above, namely the fact that each $\hat{p}_{ij}(x)$ necessarily belongs to [0, 1], which will be particularly important in the proposed testing procedure. Therefore, we will replace Assumption 2.5 by the following:

Assumption 3.1. The bandwidth sequence is such that $nh^5 \to 0$ and $nh \to \infty$.

3.3. Asymptotic distribution of the proposed test statistics under H_0

First, the asymptotic properties of the pointwise divergence criteria $\|\hat{v}(x)\|^2$ and $\hat{G}(x)$ under H_0 are stated. We have:

Lemma 3.1. *Under Assumptions 2.1–2.4 and 3.1, it holds, for any $x \in S_X^{(h)}$,*

$$\|\hat{v}(x)\|^2 \xrightarrow{\mathcal{L}} \chi^2_{(r-1)(s-1)}, \tag{3.7}$$

under H_0.

Asymptotically, $\|\hat{v}(x)\|^2$ therefore consists in a χ^2-process, which is not surprising: actually, the χ^2 limit law of the Pearson's test statistic in a classical chi-square test is implied by the asymptotic normality of the maximum likelihood estimator of π on which it is based. Here, $\|\hat{v}(x)\|^2$ is computed from $\hat{p}(x)$, also known to be asymptotically normal by (3.6). Apart from the nonparametric-rated normalisation, the situation is similar, so that the limit law of $\|\hat{v}(x)\|^2$ is logically found to be likewise $\chi^2_{(r-1)(s-1)}$ for any fixed x. However, it turns out that only the asymptotic properties of this process (expectation, variance and covariance structure) will be of importance in the further developments, and the proof of Lemma 3.1 is omitted (it can however be found in [17]). We rather explicitly derive the features of interest.

Lemma 3.2. *Under Assumptions 2.1–2.4 and 3.1, for any $x \in S_X^{(h)}$, it holds, under H_0,*

$$\mathbb{E}(\|\hat{v}(x)\|^2) = (r-1)(s-1) + O((nh)^{-1/2}) + O(nh^5)$$
$$\mathbb{V}\mathrm{ar}(\|\hat{v}(x)\|^2) = 2(r-1)(s-1) + O((nh)^{-1/2}) + O(nh^5)$$

and for any $x_1, x_2 \in S_X^{(h)}$,

$$\mathbb{C}\mathrm{ov}(\|\hat{v}(x_1)\|^2, \|\hat{v}(x_2)\|^2) = 2(r-1)(s-1)\left(\frac{\nu_0(\delta)}{\nu_0}\right)^2 (1 + O((nh)^{-1/2}) + O(nh^5)), \tag{3.8}$$

as $n \to \infty$, with $\delta = \frac{x_1 - x_2}{h}$ and $\nu_0(\delta)$ as in Lemma 2.1. Besides, the order of the remainder terms holds uniformly in x, x_1 and x_2 in $S_X^{(h)}$.

Proof. See Appendix. □

Now, to deal with expression (3.5), remind that in the classical unconditional case, the Likelihood Ratio test statistic and the chi-square test statistic are asymptotically equivalent under the null. We show below that this remains true for $\|\hat{v}(x)\|^2$ and $\hat{G}(x)$, for any $x \in S_X^h$. Indeed, let

$$\Delta_{ij}(x) = \frac{\hat{p}_{ij}(x) - \hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)},$$

and see that (3.5) can be written

$$\hat{G}(x) = 2\frac{nh\hat{f}(x)}{\nu_0}\sum_{ij}\hat{p}_{ij}(x)\log\left(1+\Delta_{ij}(x)\right).$$

Clearly, under H_0, $\sum_{ij}\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)\Delta_{ij}(x) = 0$ and $\Delta_{ij}(x) = O_P((nh)^{-1/2})$ uniformly in x under the assumed regularity conditions, so that it follows

$$\begin{aligned}
\hat{G}(x) &= 2\frac{nh\hat{f}(x)}{\nu_0}\sum_{ij}\left(\hat{p}_{ij}(x) - \hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) + \hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)\right)\left(\Delta_{ij}(x) - \frac{1}{2}\Delta_{ij}^2(x) + O_P((nh)^{-3/2})\right)\\
&= 2\frac{nh\hat{f}(x)}{\nu_0}\sum_{ij}\left(\left(\hat{p}_{ij}(x) - \hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)\right)\Delta_{ij}(x) - \frac{1}{2}\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)\Delta_{ij}^2(x) + O_P((nh)^{-3/2})\right)\\
&= \frac{nh\hat{f}(x)}{\nu_0}\sum_{ij}\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)\Delta_{ij}^2(x) + O_P((nh)^{-1/2})\\
&= \|\hat{v}(x)\|^2 + O_P((nh)^{-1/2}),
\end{aligned}$$

with the $O_P((nh)^{-1/2})$ term holding uniformly in x. This result will be used to easily derive the asymptotic distribution of G under H_0 from the one of V^2.

In that aim, see first of all that the random variables $\{\|\hat{v}(X_k)\|^2\}$ do certainly not form a sequence of independent observations: any $(X_{k'}, Z_{k'})$ such that $\|X_k - X_{k'}\| < h$ is used, through $\hat{p}$, in the computation of $\|\hat{v}(X_k)\|^2$. In fact, $\|\hat{v}(X_k)\|^2$ and $\|\hat{v}(X_{k'})\|^2$ for which $|X_k - X_{k'}| \leq 2h$ are partially built on a common set of observations, and there consequently exists some positive dependence between such two variables. On the other hand, if $|X_k - X_{k'}| > 2h$, $\|\hat{v}(X_k)\|^2$ and $\|\hat{v}(X_{k'})\|^2$ are independent (see also Remark 2.1). Therefore, with a reorganisation of the observations such that $X_1 \leq X_2 \leq \cdots \leq X_n$, $\{\|\hat{v}(X_k)\|^2\}_{k=1}^n$ asymptotically forms a sequence of m-dependent variables, with $m \doteq m_n$ growing with the sample size. Indeed, for a fixed X_k, the cardinality of the set $\{k' : \|X_k - X_{k'}\| < 2h\}$ tends in probability to $4nhf(X_k)$. As f is assumed uniformly bounded on S_X, it is sufficient to take m_n such that

$$\frac{m_n}{nh} \to 4\|f\|_\infty. \tag{3.9}$$

Now, define[2] $\nu_0(u) = (K * K)(u)$ as in Section 2, $N_0 = \int \nu_0^2(u)\mathrm{d}u$ and $\phi_0 = \int f^2(x)\mathrm{d}x$. The first two moments of the test statistic under H_0 are provided by the following result.

Lemma 3.3. *Under Assumptions 2.1–2.4, if $h \sim n^{-\beta}$ with $\beta \in]2/9, 1/2[$, it holds, under H_0,*

$$\mathbb{E}(V^2) = (r-1)(s-1) + o(h^{1/2})$$

and

$$\mathbb{V}\mathrm{ar}(V^2) = 2h(r-1)(s-1)\frac{\phi_0 N_0}{\nu_0^2} + o(h),$$

as $n \to \infty$.

Proof. See Appendix. □

The main result of this paper can now be stated.

Theorem 3.1. *Under Assumptions 2.1–2.4, if $h \sim n^{-\beta}$ with $\beta \in]2/9, 1/2[$, it holds*

$$h^{-1/2}(V^2 - (r-1)(s-1)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{2\phi_0 N_0 (r-1)(s-1)}{\nu_0^2}\right)$$

under H_0.

[2] Note that N_0 and ν_0 are constants depending on the kernel only. For example, $N_0 = 0.4337945$ and $\nu_0 = 0.6$ for the Epanechnikov kernel $K(u) = (1-u^2)\mathbb{1}(|u| \leq 1)$.

Proof. See Appendix. □

The limit law of G is directly deduced from this last result. Indeed, as

$$\begin{aligned} G &= \frac{1}{n}\sum_{k=1}^{n} \hat{G}(X_k)\mathbb{1}_{\{X_k \in S_X^h\}} \\ &= \frac{1}{n}\sum_{k=1}^{n} \left(\|\hat{v}(X_k)\|^2 + O_P((nh)^{-1/2})\right)\mathbb{1}_{\{X_k \in S_X^h\}} \\ &= V^2 + O_P((nh)^{-1/2}), \end{aligned}$$

that is $h^{-1/2}(G - V^2) = O_P(n^{-1/2}h^{-1}) = o_p(1)$ if $nh^2 \to \infty$. Therefore, it follows:

Theorem 3.2. *If Assumptions 2.1–2.4 hold, if $h \sim n^{-\beta}$ for $\beta \in]2/9, 1/2[$, then, under H_0,*

$$h^{-1/2}(G - (r-1)(s-1)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{2\phi_0 N_0 (r-1)(s-1)}{\nu_0^2}\right).$$

Now, for a practical use of the results of Theorems 3.1 and 3.2, the variance of V^2 or G has to be estimated, as it depends on the unknown coefficient $\phi_0 = \int f^2(x)\mathrm{d}x$. This can be done via the estimation of ϕ_0 by

$$\hat{\phi}_0 = \frac{1}{n}\sum_{k} \hat{f}(X_k),$$

where $\hat{f}(x)$ is the nonparametric estimator (3.3) of the density. The limit law of the normalised test statistic will not be affected by this estimation, as $|\hat{\phi}_0 - \phi_0| = O_P((nh)^{-1/2})$, by standard results on kernel density estimation. The difference between the values of the test statistic normalised with ϕ_0 and with $\hat{\phi}_0$ is therefore of order $O_P(n^{-1/2}h^{-1})$, which is $o_P(1)$ with the assumed restriction on the bandwidth $h \sim n^{-\beta}$ with $\beta < 1/2$.

Finally, two rejection criteria for the null hypothesis (1.2) at asymptotic level α are given by

$$V^2 > (r-1)(s-1) + z_{1-\alpha}\frac{\sqrt{h}}{\nu_0}\sqrt{2(r-1)(s-1)\hat{\phi}_0 N_0} \tag{3.10}$$

and

$$G > (r-1)(s-1) + z_{1-\alpha}\frac{\sqrt{h}}{\nu_0}\sqrt{2(r-1)(s-1)\hat{\phi}_0 N_0},$$

with $z_{1-\alpha}$ the $(1-\alpha)$-quantile of the standard normal distribution. Note that as in the classical χ^2-test, the test is unilateral, for evident reasons.

3.4. Consistency of the test

The consistency of the above procedures can readily be proved. Consider the pointwise divergence criterion $\|\hat{v}(x)\|^2$. The essential difference under the alternative hypothesis is the presence of an unbounded term

$$\frac{nh\hat{f}(x)}{\nu_0}\sum_{ij=11}^{rs}\frac{\left(\pi_{ij}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)^2}{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}$$

in its development, while it was trivially zero under the null. Then, it easily follows that $\mathbb{E}(\|\hat{v}(x)\|^2) = O(nh)$ and $\mathbb{V}\mathrm{ar}(\|\hat{v}(x)\|^2) = O(nh)$ for any $x \in S_X^h$, therefore $\mathbb{E}(V^2) = O(nh)$ and, by arguments similar to those of Lemma 3.3, $\mathbb{V}\mathrm{ar}(V^2) = O(nh^2)$.

Now, write C_1 for $(r-1)(s-1)$ and $\tilde{V}^2 = V^2 - z_{1-\alpha}\sqrt{h}\frac{1}{\nu_0}\sqrt{2(r-1)(s-1)\hat{\phi}_0 N_0}$. The probability to reject H_0 by criterion (3.10) is given by

$$\mathbb{P}(\tilde{V}^2 > C_1) = 1 - \mathbb{P}(\tilde{V}^2 < C_1).$$

But, under H_1, we have $E(\tilde{V}^2) - C_1 = O(nh)$ and $\mathbb{V}\mathrm{ar}(\tilde{V}^2) = O(nh^2)$ so that a necessary condition to have $\tilde{V}^2 < C_1$ is $|\tilde{V}^2 - E(\tilde{V}^2)| > r_n$, for some $r_n = O(nh)$. By the Chebyshev inequality, we have also that for any $\lambda > 0$,

$$\mathbb{P}(|\tilde{V}^2 - E(\tilde{V}^2)| > \lambda) \leq \frac{1}{\lambda^2}\mathbb{V}\mathrm{ar}(\tilde{V}^2).$$

Take $\lambda = r_n$ and see that

$$\mathbb{P}(|\tilde{V}^2 - E(\tilde{V}^2)| > r_n) = O(n^{-1}),$$

so that

$$\mathbb{P}(\tilde{V}^2 > C_1) \geq 1 - O(n^{-1}).$$

The probability to reject H_0 under H_1 tends to one.

3.5. Strength of the association

As in a classical χ^2-test, the proposed criteria are limited in testing for the null hypothesis of conditional independence, but are not able to quantify the magnitude of the association once H_0 is rejected. Indeed, the observed value of V^2 or G depends on the sample size, on the bandwidth and on the size of the table, and cannot be used to compare the strength of the association in different tables, for example. In the classical case, several measures of association have been proposed in that aim, among others the Cramer's ϕ coefficient directly derived from the χ^2-criterion, the λ coefficient, measuring the improvement of the ability to predict one variable once the other is known, or the well-known odds ratio R, in the particular case of a 2×2-table. The coefficients ϕ and λ range between 0 (independence) and 1 (perfect association), while R lies between 0 and $+\infty$, the independence being characterised by $R = 1$. See a.o. [2] for more details. Actually, none of those coefficients has really obtained unanimous agreement, but it is well out of the scope of this paper to discuss that, and we only focus on how such coefficients can be generalised to the considered conditional case. The idea is the same as for the test statistics, i.e. first checking a pointwise association coefficient, and then integrating it on the domain of interest. The extension of the coefficients is straightforward from the ideas presented in the previous sections, and we define the pointwise Cramer's coefficient

$$\hat{\phi}(x) = \sqrt{\frac{\|\hat{v}(x)\|^2}{nh \min(r-1, s-1)}},$$

the pointwise predictive coefficient

$$\hat{\lambda}(x) = \frac{\sum_i \max_j \hat{p}_{ij}(x) + \sum_j \max_i \hat{p}_{ij}(x) - \max_j \hat{p}_{\cdot j}(x) - \max_i \hat{p}_{i\cdot}(x)}{2 - \max_j \hat{p}_{\cdot j}(x) - \max_i \hat{p}_{i\cdot}(x)}$$

and the pointwise odds ratio

$$\hat{R}(x) = \frac{\hat{p}_{11}(x)\hat{p}_{22}(x)}{\hat{p}_{21}(x)\hat{p}_{12}(x)}.$$

Nevertheless, this last coefficient is hard to interpret, since it is asymmetric around 1, and it seems preferable to use the absolute value of its logarithm

$$\hat{r}(x) = |\log \hat{R}(x)|.$$

See that $\hat{r}(x)$ ranges from 0 to $+\infty$, with 0 characterising the conditional independence given $X = x$. Then, the global coefficients are computed by taking the average at the observations, that is

$$\hat{\phi} = \frac{1}{n}\sum_k \hat{\phi}(X_k) \qquad \hat{\lambda} = \frac{1}{n}\sum_k \hat{\lambda}(X_k) \qquad \hat{r} = \frac{1}{n}\sum_k \hat{r}(X_k),$$

with the same interpretation of the observed values as above, in terms of conditional independence. Naturally, some inference tools should be developed to effectively use those coefficients. Also, note that analysing the functions $\hat{\phi}(x)$, $\hat{\lambda}(x)$ or $\hat{r}(x)$ themselves can also be of interest, as it would allow us to clearly identify which values of X bring important contribution to the conditional association, and which ones do not.

4. Independence versus conditional independence

It is interesting to understand what the hypotheses of independence (1.1) and conditional independence (1.2) really imply, and how they are related to each other. Denote those two hypotheses as

$$H_0^u : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall (i,j)$$

and

$$H_0^c : \pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) \quad \forall x \in S_X,\ \forall (i,j),$$

and their respective general alternatives H_1^u and H_1^c. Note that $\pi_{ij} = \mathbb{E}(\pi_{ij}(X))$ by the law of iterated expectations, and similarly for the marginal probabilities. As it will readily be seen below, the key condition linking unconditional and conditional independence is the non-correlation between the random variables $\pi_{i\cdot}(X)$ and $\pi_{\cdot j}(X)$, for any pair (i, j), which can be intuitively understood. We therefore define

$$C : \mathbb{E}(\pi_{i\cdot}(X)\pi_{\cdot j}(X)) = \mathbb{E}(\pi_{i\cdot}(X))\mathbb{E}(\pi_{\cdot j}(X)) \quad \forall (i, j),$$

which characterises this non-correlation. Suppose now that H_0^c holds. Then, we have

$$\pi_{ij} = \mathbb{E}(\pi_{ij}(X)) = \mathbb{E}(\pi_{i\cdot}(X)\pi_{\cdot j}(X)) \quad \forall (i, j).$$

Clearly, if C is true, we get

$$\pi_{ij} = \mathbb{E}(\pi_{i\cdot}(X))\mathbb{E}(\pi_{\cdot j}(X)) = \pi_{i\cdot}\pi_{\cdot j},$$

that is H_0^u. On the other hand, if condition C does not hold,

$$\exists (i, j) : \pi_{ij} \neq \mathbb{E}(\pi_{i\cdot}(X))\mathbb{E}(\pi_{\cdot j}(X)) = \pi_{i\cdot}\pi_{\cdot j},$$

that is H_1^u. With set theory notations, this amounts to

$$H_0^c \cap C \subseteq H_0^u \tag{4.1}$$

$$H_0^c \cap \bar{C} \subseteq H_1^u, \tag{4.2}$$

where $\bar{C}$ is the negation (or the complement) of C. Taking the negation of these two relations yields $H_1^c \cup \bar{C} \supseteq H_1^u$ and $H_1^c \cup C \supseteq H_0^u$, and by considering only the intersection with C and $\bar{C}$ respectively, it follows

$$H_1^u \cap C \subseteq H_1^c$$

$$H_0^u \cap \bar{C} \subseteq H_1^c. \tag{4.3}$$

Combining those relations also gives

$$H_0^c \cap H_0^u \subseteq C$$

$$H_0^c \cap H_1^u \subseteq \bar{C}, \tag{4.4}$$

which provides some interesting possible interpretations when comparing the results of the conditional and unconditional tests. For instance, (4.4) tells that if no significant association can be emphasised by the conditional test while some was by the unconditional test, that means that the predictors are related to both categorical variables marginally (this is indeed a necessary condition for $\bar{C}$), suggesting that the observed unconditional association between R and S was actually artificially implied by their respective marginal links to the predictors. Also, (4.3) informs that, even if there is no basic association between R and S, some conditional association could be detected, in particular when the predictors are marginally strongly related to R and S. Those were the two situations mentioned in the introduction and motivating the use of conditional tests.

5. A simulation study

In this section we check the small sample performance of the proposed procedures through three simulated models, all characterised for simplicity by

$$r = s = 2, \qquad p = 1$$

$$X \sim U_{[-2,2]}$$

$$\pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x) + \gamma\delta(x) \quad \forall (i, j)$$

where $\delta(x)$ is basically the greatest 3 times continuously differentiable deviation we can add to $\pi_{i\cdot}(x)\pi_{\cdot j}(x)$ in order to keep the resulting $\pi_{ij}(x)$ between 0 and 1.[3] The coefficient γ was taken equal to 0 (conditional independence), 0.1, 0.3, 0.5 and 1.[4] We considered:

$$\text{Model 1 : } \pi_{1\cdot}(x) = \exp(-x^2), \qquad \pi_{\cdot 1}(x) = \frac{\exp(-x)}{1 + \exp(-x)}$$

$$\text{Model 2 : } \pi_{1\cdot}(x) = \exp(x/2 - 1), \qquad \pi_{\cdot 1}(x) = \exp(-x - 2)$$

$$\text{Model 3 : } \pi_{1\cdot}(x) = 0.25, \qquad \pi_{\cdot 1}(x) = 0.4.$$

For each model, we generated 500 Monte Carlo replications for each sample size $n = 50$, $n = 100$, $n = 500$ and $n = 1000$. In the procedure, we used the Epanechnikov kernel, and the bandwidth was selected the following way: we first used the plug-in method proposed in [25], slightly modified to fit with (2.7), to get an estimate of the asymptotic optimal

[3] $\delta(x)$ is actually a smoothed version of $\min(1 - \pi_{1\cdot}(x)\pi_{1\cdot}(x), \pi_{2\cdot}(x)\pi_{\cdot 1}(x), \pi_{1\cdot}(x)\pi_{\cdot 2}(x), 1 - \pi_{2\cdot}(x)\pi_{\cdot 2}(x))$.

[4] Except in Model 3, as $\gamma = 1$ leads to $\pi_{1\cdot}(x) \equiv 0$. Then, we rather took $\gamma = 0.95$.

Table 1
Rejection rates, Model 1.

	$\alpha = 0.05$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
$\gamma = 0$	CS	0.078	0.070	0.064	0.056
	LR	0.104	0.102	0.090	0.082
	CMH	0.044	0.030	0.072	0.056
	χ^2	0.018	0.018	0.068	0.056
$\gamma = 0.1$	CS	0.102	0.104	0.182	0.280
	LR	0.134	0.140	0.196	0.290
	CMH	0.056	0.058	0.118	0.152
	χ^2	0.038	0.040	0.096	0.138
$\gamma = 0.3$	CS	0.138	0.298	0.856	0.990
	LR	0.186	0.358	0.846	0.992
	CMH	0.062	0.162	0.670	0.886
	χ^2	0.050	0.126	0.556	0.828
$\gamma = 0.5$	CS	0.258	0.542	1	1
	LR	0.338	0.602	1	1
	CMH	0.176	0.376	0.964	0.998
	χ^2	0.140	0.290	0.930	0.996
$\gamma = 1$	CS	0.860	0.998	1	1
	LR	0.890	1	1	1
	CMH	0.674	0.950	1	1
	χ^2	0.534	0.866	1	1

Table 2
Rejection rates, Model 2.

	$\alpha = 0.05$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
$\gamma = 0$	CS	0.074	0.070	0.068	0.052
	LR	0.110	0.104	0.100	0.098
	CMH	0.038	0.076	0.238	0.456
	χ^2	0.208	0.532	1	1
$\gamma = 0.1$	CS	0.078	0.092	0.136	0.236
	LR	0.118	0.110	0.152	0.246
	CMH	0.032	0.026	0.056	0.098
	χ^2	0.170	0.372	0.990	1
$\gamma = 0.3$	CS	0.196	0.236	0.806	0.974
	LR	0.214	0.274	0.812	0.968
	CMH	0.044	0.066	0.228	0.422
	χ^2	0.050	0.180	0.790	0.984
$\gamma = 0.5$	CS	0.326	0.512	0.996	1
	LR	0.378	0.576	0.998	1
	CMH	0.110	0.224	0.866	0.988
	χ^2	0.034	0.078	0.252	0.524
$\gamma = 1$	CS	0.840	0.996	1	1
	LR	0.866	0.996	1	1
	CMH	0.594	0.926	1	1
	χ^2	0.038	0.086	0.488	0.778

parameter, and then multiplied it by $n^{-1/20}$ to get a bandwidth of order $O(n^{-1/4})$ (undersmoothing). The null hypothesis was rejected with a nominal level $\alpha = 0.05$. The observed percentages of rejection, for the nonparametric conditional chi-squared (CS) and likelihood ratio (LR) tests, are shown in Tables 1–3. Also, our methodologies were compared to the exact Cochran–Mantel–Haenszel test (CMH), after that the covariate X was discretised in two groups ($X < 0$ and $X \geq 0$). For information, the rejection rate of the classical χ^2-test performed on the generated contingency tables, that is totally ignoring the covariate information, is also reported.

For the three models, we observe that the CS procedure provides very satisfactory results. Small sample sizes already lead to good level, and the test seems able to detect even small deviations from the null, once n is large enough. On the other hand, under H_0, the convergence of G (LR test) toward its limit law seems slower. Interestingly enough, this observation is also proper when comparing Pearson's chi-square to the likelihood ratio test in the classical framework. Now, checking the behaviour of the classical χ^2 is also interesting in the light of the comments made in the previous section. For Model 1, it can be seen that the correlation between $\pi_{1\cdot}(X)$ and $\pi_{\cdot 1}(X)$ is zero. Therefore, (4.1) implies that the hypothesis of independence also holds when $\gamma = 0$, which explains the observed low rejection rates for H_0^u. On the other hand, for Model 2, elementary calculations show that the correlation between $\pi_{1\cdot}(X)$ and $\pi_{\cdot 1}(X)$ is high ($\simeq$0.8). When $\gamma = 0$, the clear rejection of H_0^u is thus logical in view of (4.2). For Model 3, as the considered conditional probabilities actually do not depend on X, the tested hypotheses of conditional and unconditional independence are equivalent. Interestingly, the performances of CS

Table 3
Rejection rates, Model 3.

	$\alpha = 0.05$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$
$\gamma = 0$	CS	0.080	0.078	0.072	0.068
	LR	0.122	0.116	0.084	0.082
	CMH	0.024	0.030	0.034	0.056
	χ^2	0.020	0.022	0.032	0.052
$\gamma = 0.1$	CS	0.110	0.114	0.284	0.456
	LR	0.138	0.158	0.280	0.448
	CMH	0.068	0.072	0.352	0.582
	χ^2	0.050	0.060	0.328	0.578
$\gamma = 0.3$	CS	0.318	0.470	0.984	1
	LR	0.370	0.498	0.984	1
	CMH	0.238	0.496	1	1
	χ^2	0.192	0.458	1	1
$\gamma = 0.5$	CS	0.580	0.880	1	1
	LR	0.616	0.880	1	1
	CMH	0.638	0.936	1	1
	χ^2	0.582	0.928	1	1
$\gamma = 0.95$	CS	0.986	1	1	1
	LR	0.978	1	1	1
	CMH	1	1	1	1
	χ^2	1	1	1	1

Table 4
Initial data.

		Claims		
		No	Yes	
Sex	M	8 324	1034	9 358
	F	4 638	509	5 147
		12 962	1543	14 505

and χ^2 are found to be similar, so that even if one suspects that X has absolutely no effect on R and S, nothing is lost, in terms of level or power, when working conditionally on it. Finally, the CMH test, competitor of CS and LR when testing for the conditional independence, appears to behave erratically. When the unconditional and conditional independence hypotheses are in accord (Model 1 and Model 3), this test performs as expected, that is correctly but still with some lost of power compared to CS and LR, due to the discretisation of X. However, in Model 2, it seems a bit confused between the disagreeing independence concepts, which yields poor performance. Obviously, this behaviour could be explained and quantified by carefully looking at the conditional odds ratio in each of the defined strata, with our particular choice of link functions and design. Anyway, focusing on CS, the results appear to be very satisfactory in a general way, all the more since in the considered case of a 2×2 table, the limit law of the $\|\hat{v}(X_k)\|^2$ under H_0 is χ^2_1, that is the more asymmetric χ^2 distribution. The convergence toward the normal limit distribution under the null hypothesis is thus probably slower than in any other situation.

6. Real data example

We illustrate our method with its application on a real data set, which consists on a portfolio of 14 505 car insurance policies of a Belgian insurance company. For each insured are known the sex and the number of notified claims during one year (1997). A question of interest for the company is to know if there is a relationship between the sex and the notification of claims, in order to efficiently adjust risk-rated premiums. The data, presented as a contingency table crossing the variable sex with the variable claims, are shown in Table 4.

The classical chi-square test on these data yields $U^2 = 4.58$, $df = 1$, p-value $= 0.03$, so that the hypothesis of independence between sex and claims is rejected at the level 0.05. The sign of the residuals with respect to the independence model besides indicates that men do notify claims more often than women. To go further in the analysis of this association, we would like to identify the factors possibly responsible for it. For example, it turns out that, for each insured, we also know the power of the insured vehicle (in kW). The power of the vehicle is possibly related to the sex (men maybe drive more powerful cars than women) and to the number of notified accidents as well (the faster you go, the more often you crash). The le Cessie-van Houwelingen goodness-of-fit test [26] rejects (p-value $= 0.0015$) the marginal logistic model for the link between power and sex, so that a model like (2.3) has already to be discarded. Instead, we compute the nonparametric estimator (2.6). The joint and the marginal conditional probabilities are represented in Fig. 1. The effect of the power on both sex and claims seems clear: up to what can be considered as some boundary effect (power > 200), the marginal

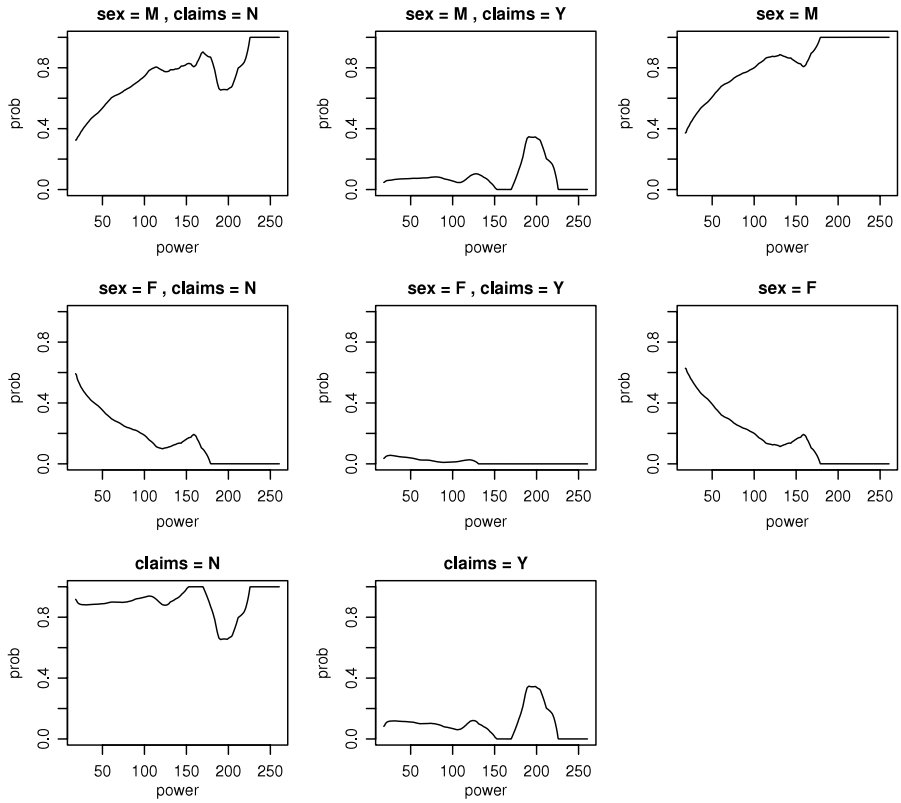

Fig. 1. Nonparametric estimation of the joint and marginal conditional distributions.

probabilities are as expected. Therefore, in order to control for this effect when analysing the association of interest, we test the independence between sex and claims, given the power of the vehicle, by our procedure. Using the Epanechnikov kernel and a bandwidth selected in the same way as in the simulation study (actual value: 6.77), one finds a p-value of 0.21, so that the conditional independence hypothesis is not rejected. This means that if a man and a woman are driving a vehicle of equal power, there is no evidence that the man is more exposed to the risk than the woman, contrary to what was observed when the power was not considered. In other words, from the observations made in Section 4, it is likely that the power of the driven car is a factor inducing the dependence between sex and claims emphasised by the classical χ^2-test.

7. Conclusion and perspectives

This paper addresses the problem of testing for the independence between two categorical variables R and S, given a vector X of continuous covariates. This kind of procedure is particularly useful when it is known that some own characteristics of each individual may influence the two categorical variables, as well as their possible relationship. Indeed, working conditionally on these covariates allows us to take into account, and to control in some sense, their effect on the association observed in the contingency table. The test procedure is at a first step based on a pointwise divergence criterion, basically some generalisation of the classical chi-square or the likelihood ratio criteria, between the estimated joint conditional distribution of R and S and the product of their estimated conditional marginal distributions. These conditional distributions, regarded as regression functions, are nonparametrically estimated by Nadaraya–Watson-like estimators. This is seen to be particularly well adapted to the setting of conditional probabilities (easy and fast computation, structural properties of the true probabilities maintained for the estimates), without the need of any hazardous parametric assumption. Then, the pointwise divergence is integrated by computing its empirical mean at the observations, which provides the test statistic. This one is shown to asymptotically follow a normal distribution under the conditional independence hypothesis, so that a practical rejection criterion can be derived. Simulations show the good behaviour of this asymptotic procedure in finite sample situations. As far as we know, the problem of directly testing for an hypothesis such as (1.2) in contingency tables is new, and the results presented in this paper are promising. However, the proposed procedures are likely to be improved or discussed in various ways. Some bootstrap algorithms could possibly improve the performance of the asymptotic rejection criteria. Also, it seems worth thinking about a way to select the bandwidth maximising the power of the resulting test. These ideas are well out of the scope of this paper, and are let as open questions.

Acknowledgments

Research support from the "Interuniversity Attraction Pole", Phase VI (No. P06/03) from the Belgian Science Policy, is acknowledged. This work was also partially supported by a FSR grant from the Université catholique de Louvain, and a Centre of Excellence grant to MASCOS from the Australian Research Council. Finally, Pr. I. Van Keilegom is gratefully acknowledged for useful discussions.

Appendix

Some preliminary technical lemmas are first expounded. Then, the proofs of the main results stated in the paper are provided.

Lemma A.1. *Under Assumptions 2.1–2.4, if $h = o(n^{-1/5})$, we have, for any $x \in S_X^h$, for any (i, j) and for any positive integer α,*

$$\mathbb{E}\left(\left(\hat{p}_{ij}(x) - \pi_{ij}(x)\right)^{2\alpha}\right) = \frac{\nu_0^\alpha}{(nh)^\alpha f^\alpha(x)} \frac{(2\alpha)!}{\alpha! 2^\alpha} \left(\pi_{ij}(x)(1-\pi_{ij}(x))^\alpha\right) + o((nh)^{-\alpha})$$

as $n \to \infty$, and the order of the remainder term holds uniformly in $x \in S_X^h$.

Proof. In the context induced by Assumption 2.1, the residuals are bounded in absolute value by 1, and their conditional distribution given X depends only on π_{ij}. Therefore, the result is the direct application of Theorem 1.3.3 and Corollary 1.3.3 of [17]. See also [27]. □

Corollary A.1. *Under Assumptions 2.1–2.4, if $h = o(n^{-1/5})$, for any $x \in S_X^h$, for any (i, j), for any positive integer α, we have, as $n \to \infty$,*

(i) $\mathbb{E}((\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x))^{2\alpha}) = O((nh)^{-\alpha})$;
(ii) $\mathbb{E}((\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x))^{2\alpha}) = O((nh)^{-\alpha})$;
(iii) $\mathbb{E}((\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x))^{2\alpha}(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x))^{2\alpha}) = O((nh)^{-2\alpha})$;
(iv) $\mathbb{E}((\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x))^{2\alpha}) = O((nh)^{-\alpha})$;
(v) $\mathbb{E}((\hat{f}(x) - f(x))^{2\alpha}) = O((nh)^{-\alpha})$.

Besides, these orders hold uniformly in $x \in S_X^h$.

Proof. (i) and (ii) are exact copies of the result of Lemma A.1, for the marginal conditional probabilities. For (iii), use the Cauchy–Schwartz inequality to find that

$$\mathbb{E}\left(\left(\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x)\right)^{2\alpha}\left(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x)\right)^{2\alpha}\right) \le \left(\mathbb{E}\left(\left(\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x)\right)^{4\alpha}\right)\right)^{1/2}\left(\mathbb{E}\left(\left(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x)\right)^{4\alpha}\right)\right)^{1/2},$$

and use (i) and (ii). (iv) is proved seeing that

$$\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x) = (\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x))\pi_{\cdot j}(x) + (\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x))\pi_{i\cdot}(x) + (\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x))(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x)),$$

so that $\mathbb{E}((\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x))^{2\alpha}) = O(\mathbb{E}((\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x)))^{2\alpha})$. Finally, (v) is proved similarly to Theorem 1.3.3 in [17], from

$$\left(\hat{f}(x) - f(x)\right)^{2\alpha} = (nh)^{-2\alpha}\left(\sum_k K\left(\frac{x - X_k}{h}\right) - hf(x)\right)^{2\alpha}. \quad \square$$

Corollary A.2. *Under Assumptions 2.1–2.4 and 2.5, if $h = o(n^{-1/5})$, for any $x \in S_X^h$, for any (i, j) and (i', j'), we have, as $n \to \infty$,*

(i) $\mathbb{E}\left(\left(\hat{p}_{ij}(x) - \pi_{ij}(x)\right)\left(\hat{p}_{i'j'}(x) - \pi_{i'j'}(x)\right)\right) = \dfrac{\nu_0}{(nh)f(x)}\pi_{i'j'}(x)(\delta_{i,i'}\delta_{j,j'} - \pi_{ij}(x)) + O(n^{-1}h) + O(h^4)$

(ii) $\mathbb{C}\mathrm{ov}\left(\left(\hat{p}_{ij}(x) - \pi_{ij}(x)\right)^2, \left(\hat{p}_{i'j'}(x) - \pi_{i'j'}(x)\right)^2\right)$

$$= \frac{2\nu_0^2}{(nh)^2 f^2(x)}\left(\pi_{i'j'}(x)(\delta_{i,i'}\delta_{j,j'} - \pi_{ij}(x))\right)^2 + O((nh)^{-3}) + O(n^{-1}h^3)$$

(iii) $\mathbb{E}\left(\left(\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x)\right)\left(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x)\right)\right) = O(n^{-1}h) + O(h^4)$ *under (1.2).*

Besides, the order of the remainder terms holds uniformly in $x \in S_X^h$.

Proof. Results (i) and (ii) are very similar to Lemma A.1, and their proof is also largely inspired from the one of Theorem 1.3.3 in [17]. See [17, Section 5.3] for details. The proof of (iii) is trivial, by writing

$$\mathbb{E}\left(\left(\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x)\right)\left(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x)\right)\right) = \sum_{i'}\sum_{j'}\mathbb{E}\left(\left(\hat{p}_{ij'}(x) - \pi_{ij'}(x)\right)\left(\hat{p}_{i'j}(x) - \pi_{i'j}(x)\right)\right),$$

and replacing by expressions found in (i), given that $\pi_{ij}(x) = \pi_{i\cdot}(x)\pi_{\cdot j}(x)$. □

Proof of Lemma 2.1

Again, the proof is an adaptation of the proof of Theorem 1.3.3 in [17], so that we hereafter present only a sketch of it. We derive only (2.12) in the case $(i,j)=(i',j')$. The other situations can be handled *mutatis mutandis*. We have

$$\mathbb{C}\mathrm{ov}(\hat{p}_{ij}(x_1),\hat{p}_{ij}(x_2)) = \mathbb{E}((\hat{p}_{ij}(x_1)-\pi_{ij}(x_1))(\hat{p}_{ij}(x_2)-\pi_{ij}(x_2))) - \mathbb{E}(\hat{p}_{ij}(x_1)-\pi_{ij}(x_1))\mathbb{E}(\hat{p}_{ij}(x_2)-\pi_{ij}(x_2)). \tag{A.1}$$

The first term in the right-hand size can be developed as

$$\begin{aligned}
&\mathbb{E}((\hat{p}_{ij}(x_1)-\pi_{ij}(x_1))(\hat{p}_{ij}(x_2)-\pi_{ij}(x_2))) \\
&\quad = \frac{(nh)^{-2}\left(\sum_k K\left(\frac{x_1-X_k}{h}\right)(Z_k^{(ij)}-\pi_{ij}(x_1))\right)\left(\sum_k K\left(\frac{x_2-X_k}{h}\right)(Z_k^{(ij)}-\pi_{ij}(x_2))\right)}{(nh)^{-2}\left(\sum_k K\left(\frac{x_1-X_k}{h}\right)\right)\left(\sum_k K\left(\frac{x_2-X_k}{h}\right)\right)} \\
&\quad \doteq \frac{\hat{N}}{\hat{D}}.
\end{aligned}$$

Write the numerator as

$$\begin{aligned}
\hat{N} = (nh)^{-2}\Bigg(&\sum_k K\left(\frac{x_1-X_k}{h}\right)K\left(\frac{x_2-X_k}{h}\right)(Z_k^{(ij)}-\pi_{ij}(x_1))(Z_k^{(ij)}-\pi_{ij}(x_2)) \\
&+\sum_k\sum_{k'\neq k} K\left(\frac{x_1-X_k}{h}\right)K\left(\frac{x_2-X_{k'}}{h}\right)(Z_k^{(ij)}-\pi_{ij}(x_1))(Z_{k'}^{(ij)}-\pi_{ij}(x_2))\Bigg),
\end{aligned}$$

and see that the second term will simplify with $\mathbb{E}(\hat{p}_{ij}(x_1)-\pi_{ij}(x_1))\mathbb{E}(\hat{p}_{ij}(x_2)-\pi_{ij}(x_2))$ in (A.1) up to a uniformly $O(n^{-1})$ term, so that we take the liberty to omit it in the sequel. As we have

$$\mathbb{E}((Z_k^{(ij)}-\pi_{ij}(x_1))(Z_k^{(ij)}-\pi_{ij}(x_2))|X_k) = \pi_{ij}(X_k)(1-\pi_{ij}(X_k)) + (\pi_{ij}(X_k)-\pi_{ij}(x_1))(\pi_{ij}(X_k)-\pi_{ij}(x_2))$$

and $\mathbb{E}((Z_k^{(ij)}-\pi_{ij}(x))|X_k) = \pi_{ij}(X_k)-\pi_{ij}(x)$, we can write

$$\begin{aligned}
\mathbb{E}(\hat{N}) &= \mathbb{E}(\mathbb{E}(\hat{N}|\{X_k\})) \\
&= (nh)^{-2}\left[n\mathbb{E}\left(K\left(\frac{x_1-X}{h}\right)K\left(\frac{x_2-X}{h}\right)\pi_{ij}(X)(1-\pi_{ij}(X))\right)\right. \\
&\quad \left. + n\mathbb{E}\left(K\left(\frac{x_1-X}{h}\right)K\left(\frac{x_2-X}{h}\right)(\pi_{ij}(X)-\pi_{ij}(x_1))(\pi_{ij}(X)-\pi_{ij}(x_2))\right)+\cdots\right].
\end{aligned}$$

The first expectation in the bracket can be computed as

$$\begin{aligned}
&\int K\left(\frac{x_1-y}{h}\right)K\left(\frac{x_2-y}{h}\right)\pi_{ij}(y)(1-\pi_{ij}(y))f(y)\mathrm{d}y \\
&= h\int K(u)K(u+\delta)\pi_{ij}(x_1-uh)(1-\pi_{ij}(x_1-uh))f(x_1-uh)\mathrm{d}u \\
&= h\pi_{ij}(x_1)(1-\pi_{ij}(x_1))f(x_1)\int K(u)K(u+\delta)\mathrm{d}u + O(h^2)\times\int uK(u)K(u+\delta)\mathrm{d}u \\
&= h\pi_{ij}(x_1)(1-\pi_{ij}(x_1))f(x_1)\nu_0(\delta)(1+O(h)),
\end{aligned}$$

with the change of variable $u=(x_1-x_2)/h$. See also that as the $O(h)$ term essentially consists of products of the functions π_{ij}, f and their derivatives, Assumptions 2.2 and 2.3 ensure that it holds uniformly in x. Similarly, the other expectation can be shown to be uniformly $O(h^3)$, so that it remains

$$\mathbb{E}(\hat{N}) = (nh)^{-1}\pi_{ij}(x_1)(1-\pi_{ij}(x_1))f(x_1)\nu_0(\delta)(1+O(h)).$$

The same kind of development leads to

$$\mathbb{E}(\hat{D}) = f(x_1)f(x_2)+O((nh)^{-1})+O(h^2),$$

$\mathbb{C}\mathrm{ov}(\hat{N},\hat{D}) = O((nh)^{-2})$ and $\mathbb{V}\mathrm{ar}(\hat{D}) = O((nh)^{-1})$, so that, as from a Taylor expansion of $\frac{\hat{N}}{\hat{D}}$ around $\frac{\mathbb{E}(\hat{N})}{\mathbb{E}(\hat{D})}$ we can write

$$\mathbb{E}\left(\frac{\hat{N}}{\hat{D}}\right) = \frac{\mathbb{E}(\hat{N})}{\mathbb{E}(\hat{D})} + O(\mathbb{C}\mathrm{ov}(\hat{N},\hat{D}) + \mathbb{E}(\hat{N})\mathbb{V}\mathrm{ar}(\hat{D})), \tag{A.2}$$

we find

$$\mathbb{C}\mathrm{ov}(\hat{p}_{ij}(x_1), \hat{p}_{ij}(x_2)) = \frac{\pi_{ij}(x_1)(1-\pi_{ij}(x_1))\nu_0(\delta)}{nhf(x_2)}(1+O(h)).$$

Moreover, note that

$$\frac{\nu_0(\delta)}{f(x_2)} = \frac{\nu_0(\delta)(1+O(h))}{f(x_1)}$$

so that finally

$$\mathbb{C}\mathrm{ov}(\hat{p}_{ij}(x_1), \hat{p}_{ij}(x_2)) = \frac{\pi_{ij}(x_1)(1-\pi_{ij}(x_1))\nu_0(\delta)}{nhf(x_1)}(1+O(h)).$$

See that any remainder term is clearly uniformly bounded under Assumptions 2.2 and 2.3, and the proof is completed. □

Proof of Lemma 3.2

First of all the expectation and the covariance matrix of the vector $\hat{v}(x)$ are derived. Decompose any component $\hat{v}_{ij}(x)$ as

$$\begin{aligned}
\hat{v}_{ij}(x) = v_{ij}(x) &+ \frac{\sqrt{nh}}{\sqrt{\nu_0}}\frac{\sqrt{f(x)}}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}}\left(\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right) \\
&+ \frac{\sqrt{nh}}{\sqrt{\nu_0}}\sqrt{f(x)}\left(\hat{p}_{ij}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)\left(\frac{1}{\sqrt{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}} - \frac{1}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}}\right) \\
&+ \frac{\sqrt{nh}}{\sqrt{\nu_0}}\frac{\left(\hat{p}_{ij}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}}\left(\sqrt{\hat{f}(x)} - \sqrt{f(x)}\right) \\
&+ \frac{\sqrt{nh}}{\sqrt{\nu_0}}\sqrt{f(x)}\left(\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)\left(\frac{1}{\sqrt{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}} - \frac{1}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}}\right) \\
&+ \frac{\sqrt{nh}}{\sqrt{\nu_0}}\frac{\left(\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}}\left(\sqrt{\hat{f}(x)} - \sqrt{f(x)}\right) \\
&+ \frac{\sqrt{nh}}{\sqrt{\nu_0}}\left(\hat{p}_{ij}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)\left(\frac{1}{\sqrt{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}} - \frac{1}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}}\right)\left(\sqrt{\hat{f}(x)} - \sqrt{f(x)}\right) \\
&+ \frac{\sqrt{nh}}{\sqrt{\nu_0}}\left(\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)\right)\left(\frac{1}{\sqrt{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}} - \frac{1}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}}\right)\left(\sqrt{\hat{f}(x)} - \sqrt{f(x)}\right),
\end{aligned}$$

which can be written as

$$\hat{v}_{ij}(x) = v_{ij}(x) + I_{ij}(x) + II_{ij}(x) + III_{ij}(x) + IV_{ij}(x),$$

where $IV_{ij}(x)$ represents the terms 5 to 8 in the above decomposition. Defining the corresponding vectors $I(x)$, $II(x)$, $III(x)$ and $IV(x)$, we have

$$\hat{v}(x) = v(x) + I(x) + II(x) + III(x) + IV(x). \tag{A.3}$$

Now, denote

$$\begin{aligned}
\pi_0(x) &= (\pi_{1\cdot}(x), \pi_{2\cdot}(x), \ldots, \pi_{r-1\cdot}(x), \pi_{\cdot 1}(x), \ldots, \pi_{\cdot s-1}(x))^t, \\
\hat{p}_0(x) &= \left(\hat{p}_{1\cdot}(x), \hat{p}_{2\cdot}(x), \ldots, \hat{p}_{r-1\cdot}(x), \hat{p}_{\cdot 1}(x), \ldots, \hat{p}_{\cdot s-1}(x)\right)^t,
\end{aligned}$$

the set

$$\Theta = \left\{\theta \in \mathbb{R}^{r+s-2} : \theta^{(q)} > 0 \forall q, \sum_{i=1}^{r-1}\theta^{(i)} < 1, \sum_{j=1}^{s-1}\theta^{(r-1+j)} < 1\right\}$$

and the function $\Pi^0 : \Theta \to (0,1)^{rs}$, $\Pi^0(\theta) = (\Pi^0_{11}(\theta), \Pi^0_{12}(\theta), \ldots, \Pi^0_{r(s-1)}(\theta), \Pi^0_{rs}(\theta))^t$, with

$$\Pi^0_{ij}(\theta) = \begin{cases} \theta^{(i)}\theta^{(r-1+j)} & \text{if } 1 \le i \le r-1,\ 1 \le j \le s-1 \\ \theta^{(i)}\left(1 - \sum_{j'=1}^{s-1} \theta^{(r-1+j')}\right) & \text{if } 1 \le i \le r-1,\ j = s \\ \left(1 - \sum_{i'=1}^{r-1} \theta^{(i')}\right)\theta^{(r-1+j)} & \text{if } i = r,\ 1 \le j \le s-1 \\ \left(1 - \sum_{i'=1}^{r-1} \theta^{(i')}\right)\left(1 - \sum_{j'=1}^{s-1} \theta^{(r-1+j')}\right) & \text{if } i = r,\ j = s. \end{cases}$$

See that the function $\Pi^0(\theta)$ is two times continuously differentiable with respect to any $\theta^{(q)}$, with e.g.

$$\frac{\partial \Pi^0_{ij}}{\partial \theta^{(q)}}(\theta) = \begin{cases} \theta^{(r-1+j)}\delta_{i,q} + \theta^{(i)}\delta_{r-1+j,q} & \text{if } 1 \le i \le r-1,\ 1 \le j \le s-1 \\ \left(1 - \sum_{j'=1}^{s-1} \theta^{(r-1+j')}\right)\delta_{i,q} - \theta^{(i)}\left(\sum_{q'=r}^{r+s-2} \delta_{q,q'}\right) & \text{if } 1 \le i \le r-1,\ j = s \\ -\theta^{(r-1+j)}\left(\sum_{q'=1}^{r-1} \delta_{q,q'}\right) + \left(1 - \sum_{i'=1}^{r-1} \theta^{(i')}\right)\delta_{r-1+j,q} & \text{if } i = r,\ 1 \le j \le s-1 \\ -\left(\left(1 - \sum_{j'=1}^{s-1} \theta^{(r-1+j')}\right)\left(\sum_{q'=1}^{r-1} \delta_{q,q'}\right) + \left(1 - \sum_{i'=1}^{r-1} \theta^{(i')}\right)\left(\sum_{q'=r}^{r+s-2} \delta_{q,q'}\right)\right) & \text{if } i = r,\ j = s \end{cases} \tag{A.4}$$

where $\delta_{q,q'}$ denotes the Kronecker delta. From that, define the $[(rs) \times (r+s-2)]$-matrix A_x as

$$[A_x]_{ij,q} = \frac{\sqrt{f(x)}}{\sqrt{\nu_0}\sqrt{\Pi^0_{ij}(\pi_0(x))}} \frac{\partial \Pi^0_{ij}}{\partial \theta^{(q)}}(\pi_0(x)). \tag{A.5}$$

Like the already defined vectors, the rows of A_x are indexed by the pairs ij, in the same order as in vectors $\hat{p}(x)$ and $\pi(x)$, in (2.5) and (2.10). Thus, the index ij denotes the $((i-1)s+j)$th row of A_x. We have

$$I_{ij}(x) = \frac{\sqrt{nh}}{\sqrt{\nu_0}} \frac{\sqrt{f(x)}}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}} \left(- \sum_{q=1}^{r+s-2} \left(\hat{p}_0^{(q)}(x) - \pi_0^{(q)}(x)\right) \frac{\partial \Pi^0_{ij}}{\partial \theta^{(q)}}(\pi_0(x)) \right.$$
$$\left. - \frac{1}{2} \sum_{q=1}^{r+s-2} \sum_{q'=1}^{r+s-2} \left(\hat{p}_0^{(q)}(x) - \pi_0^{(q)}(x)\right)\left(\hat{p}_0^{(q')}(x) - \pi_0^{(q')}(x)\right) \frac{\partial^2 \Pi^0_{ij}}{\partial \theta^{(q)} \partial \theta^{(q')}}(\pi_0(x)) \right).$$

Tedious but straightforward calculations from (A.4) show that the first term of the right-hand side equals

$$I^{(a)}_{ij}(x) \doteq \frac{\sqrt{nh}}{\sqrt{\nu_0}} \frac{\sqrt{f(x)}}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}} \left((1 - 2\delta_{i,r})(\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x))\pi_{\cdot j}(x) + (1 - 2\delta_{j,s})(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x))\pi_{i\cdot}(x)\right)$$

and similarly the second one can be seen to be

$$I^{(b)}_{ij}(x) \doteq \frac{2\sqrt{nh}}{\sqrt{\nu_0}} \frac{\sqrt{f(x)}}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}} (\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x))(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x)).$$

Define the vector $I^{(b)}(x)$ and see that

$$v(x) + I(x) = \mathcal{A}_x v(x) + I^{(b)}(x), \tag{A.6}$$

with $\mathcal{A}_x = (I - A_x(A_x^t A_x)^{-1} A_x^t)$. Since

$$\mathbb{E}\left((\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x))(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x))\right) = o((nh)^{-1})$$

under H_0 by Corollary A.2, it follows that

$$\mathbb{E}(I^{(b)}(x)) = o((nh)^{-1/2}) \times 1_{rs} \tag{A.7}$$

with 1_{rs} being a rs-vector whose components are all equal to 1. By Corollary A.1, we have also

$$\mathbb{E}\left((\hat{p}_{i\cdot}(x) - \pi_{i\cdot}(x))^2(\hat{p}_{\cdot j}(x) - \pi_{\cdot j}(x))^2\right) = O((nh)^{-2}),$$

so that $\mathbb{E}(I^{(b)}(x)I^{(b)}(x)^t) = O((nh)^{-1}) \times 1_{rs}1_{rs}^t$, and therefore

$$\mathbb{V}\mathrm{ar}(I^{(b)}(x)) = O((nh)^{-1}) \times 1_{rs}1_{rs}^t. \tag{A.8}$$

Concerning the term $II(x)$, write the Taylor expansion

$$\frac{1}{\sqrt{\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x)}} - \frac{1}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}} = -\frac{1}{2\sqrt{(\pi_{i\cdot}(x)\pi_{\cdot j}(x))^3}}(\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x)) + \text{lower order terms},$$

where the lower order terms are all controlled in L_1 by Corollary A.1. Also, as $\mathbb{E}((\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x))^2) = O((nh)^{-1})$ and $\mathbb{E}((\hat{p}_{ij}(x) - \pi_{ij}(x))^2) = O((nh)^{-1})$, it follows, by the Cauchy–Schwartz inequality, that

$$\mathbb{E}(II(x)) = O((nh)^{-1/2}) \times 1_{rs}. \tag{A.9}$$

Similarly, again from Corollary A.1, we have $\mathbb{E}((\hat{p}_{i\cdot}(x)\hat{p}_{\cdot j}(x) - \pi_{i\cdot}(x)\pi_{\cdot j}(x))^4) = O((nh)^{-2})$ and $\mathbb{E}((\hat{p}_{ij}(x) - \pi_{ij}(x))^4) = O((nh)^{-2})$, so that $\mathbb{E}(II_{ij}^2(x)) = O((nh)^{-1})$ and therefore

$$\mathbb{V}\mathrm{ar}(II(x)) = O((nh)^{-1}) \times 1_{rs}1_{rs}^t. \tag{A.10}$$

The term $III(x)$ is treated in a very similar way, with

$$\left(\sqrt{\hat{f}(x)} - \sqrt{f(x)}\right) = \frac{1}{2\sqrt{f(x)}}(\hat{f}(x) - f(x)) + \text{lower order terms},$$

so that $\mathbb{E}\left(\left(\sqrt{\hat{f}(x)} - \sqrt{f(x)}\right)^2\right) = O(\mathbb{E}((\hat{f}(x) - f(x))^2)) = O((nh)^{-1})$, which implies by Cauchy–Schwartz

$$\mathbb{E}(III(x)) = O((nh)^{-1/2}) \times 1_{rs}. \tag{A.11}$$

Also, $\mathbb{E}\left(\left(\sqrt{\hat{f}(x)} - \sqrt{f(x)}\right)^4\right) = O((nh)^{-2})$, so that

$$\mathbb{V}\mathrm{ar}(III(x)) = O((nh)^{-1}) \times 1_{rs}1_{rs}^t. \tag{A.12}$$

Finally, see that the terms in $IV_{ij}(x)$ essentially consist of products of the previous ones, so that the same treatment as above leads to an expectation at most $O((nh)^{-1/2})$ and a variance at most $O((nh)^{-1})$.

From the above, see that we can write, from (A.3), (A.6), (A.7), (A.9) and (A.11),

$$\mathbb{E}(\hat{v}(x)) = \mathcal{A}_x\mathbb{E}(v(x)) + O((nh)^{-1/2}) \times 1_{rs}, \tag{A.13}$$

and from (A.3), (A.6), (A.8), (A.10) and (A.12), with Cauchy–Schwartz,

$$\mathbb{V}\mathrm{ar}(\hat{v}(x)) = \mathcal{A}_x\mathbb{V}\mathrm{ar}(v(x))\mathcal{A}_x^t + O((nh)^{-1/2}) \times 1_{rs}1_{rs}^t. \tag{A.14}$$

It is clear that $\mathbb{E}(v(x)) = O(\sqrt{nh^5}) \times 1_{rs}$, while, from Corollary A.2(i), $\mathbb{V}\mathrm{ar}(v(x)) = I - \sqrt{\Pi^0(\pi_0(x))}\sqrt{\Pi^0(\pi_0(x))}^t + (O(h^2) + O(nh^5)) \times 1_{rs}1_{rs}^t$. Therefore, we have, from (A.13) and (A.14) and Assumption 3.1,

$$\mathbb{E}(\hat{v}(x)) = (O(\sqrt{nh^5}) + O((nh)^{-1/2})) \times 1_{rs}$$

and

$$\mathbb{V}\mathrm{ar}(\hat{v}(x)) = \Sigma_x + (O((nh)^{-1/2}) + O(nh^5)) \times 1_{rs}1_{rs}^t,$$

with

$$\begin{aligned}\Sigma_x &= \left(I - A_x\left(A_x^tA_x\right)^{-1}A_x^t\right)\left(I - \sqrt{\Pi^0(\pi_0(x))}\sqrt{\Pi^0(\pi_0(x))}^t\right)\left(I - A_x\left(A_x^tA_x\right)^{-1}A_x^t\right)^t \\ &= I - \sqrt{\Pi^0(\pi_0(x))}\sqrt{\Pi^0(\pi_0(x))}^t - A_x\left(A_x^tA_x\right)^{-1}A_x^t,\end{aligned}$$

since $A_x^t\sqrt{\Pi^0(\pi_0(x))} = 0$. Moreover, as

$$\mathrm{trace}(\Sigma_x) = \mathrm{trace}\left(I - \sqrt{\Pi^0(\pi_0(x))}\sqrt{\Pi^0(\pi_0(x))}^t\right) - \mathrm{trace}\left(A_x\left(A_x^tA_x\right)^{-1}A_x^t\right),$$

the trace of Σ_x is $(r-1)(s-1)$, as the first trace in the above expression is easily seen to be equal to $rs-1$, as $\sum_{ij}\Pi^0(\pi_0(x)) \equiv 1$, and the second one is equal to $\mathrm{trace}(A_x^tA_x(A_x^tA_x)^{-1})$, that is $r+s-2$.

Now, by the usual formula of the expectation of quadratic forms, we have

$$\begin{aligned}\mathbb{E}(\|\hat{v}(x)\|^2) &= \mathbb{E}(\hat{v}(x)^t\hat{v}(x))\\ &= \operatorname{trace}(\mathbb{V}\mathrm{ar}(\hat{v}(x))) + \mathbb{E}(\hat{v}(x))^t\mathbb{E}(\hat{v}(x))\\ &= \operatorname{trace}(\Sigma_x) + O((nh)^{-1/2}) + O(nh^5)\\ &= (r-1)(s-1) + O((nh)^{-1/2}) + O(nh^5),\end{aligned}$$

which provides the first result of the corollary.

To derive the expression of the variance, write again

$$\hat{v}(x) = \mathcal{A}_x v(x) + R(x),$$

where $R(x)$ stands for the terms $I^{(b)}(x)$ to $IV(x)$ above. The variance of $\|\hat{v}(x)\|^2$ can therefore be written

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\hat{v}(x)^t\hat{v}(x)) &= \mathbb{V}\mathrm{ar}(v(x)^t\mathcal{A}_x^t\mathcal{A}_x v(x)) + \mathbb{V}\mathrm{ar}(v(x)^t\mathcal{A}_x^t R(x)) + \cdots\\ &\quad + \mathbb{C}\mathrm{ov}(v(x)^t\mathcal{A}_x^t\mathcal{A}_x v(x),\, v(x)^t\mathcal{A}_x^t R(x)) + \cdots,\end{aligned}$$

where the first dots are for the other variance terms, and the other ones for the other covariance terms, at most of the same order as the written ones. See that

$$\begin{aligned}\mathbb{V}\mathrm{ar}(v(x)^t\mathcal{A}_x^t R(x)) &\le \mathbb{E}((v(x)^t\mathcal{A}_x^t R(x))^2)\\ &\le \mathbb{E}(\|\mathcal{A}_x v(x)\|^2\|R(x)\|^2)\\ &\le \sqrt{\mathbb{E}(\|\mathcal{A}_x v(x)\|^4)}\sqrt{\mathbb{E}(\|R(x)\|^4)}.\end{aligned}$$

From the above arguments, it can be shown that $\mathbb{E}(\|R(x)\|^4) = O((nh)^{-2})$, so that

$$\mathbb{V}\mathrm{ar}(v(x)^t\mathcal{A}_x^t R(x)) = O((nh)^{-1}),$$

and therefore

$$\mathbb{C}\mathrm{ov}(v(x)^t\mathcal{A}_x^t\mathcal{A}_x v(x),\, v(x)^t\mathcal{A}_x^t R(x)) = O((nh)^{-1/2}).$$

Moreover, as $\mathcal{A}_x$ is symmetric and idempotent, it remains

$$\mathbb{V}\mathrm{ar}(\hat{v}(x)^t\hat{v}(x)) = \mathbb{V}\mathrm{ar}(v(x)^t\mathcal{A}_x v(x)) + O((nh)^{-1/2}). \tag{A.15}$$

Now, write

$$v(x)^t\mathcal{A}_x v(x) = \sum_{ij=11}^{rs}\sum_{i'j'=11}^{rs} v_{ij}(x)v_{i'j'}(x)[\mathcal{A}_x]_{ij,i'j'}.$$

See that, from definition of matrix $\mathcal{A}_x$, we have

$$[\mathcal{A}_x]_{i'j',ij} = \delta_{i,i'}\delta_{j,j'} - \sum_{q=1}^{r+s-2}[A_x]_{i'j',q}[(A_x^tA_x)^{-1}A_x^t]_{q,ij}.$$

From (A.4) and (A.5), some direct but tedious algebra leads to

$$[\mathcal{A}_x]_{i'j',ij} = \delta_{i,i'}\delta_{j,j'} + 2\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)\pi_{i'\cdot}(x)\pi_{\cdot j'}(x)} - \delta_{i,i'}\sqrt{\pi_{\cdot j}(x)\pi_{\cdot j'}(x)} - \delta_{j,j'}\sqrt{\pi_{i\cdot}(x)\pi_{i'\cdot}(x)}.$$

Therefore, we have

$$\begin{aligned}v(x)^t\mathcal{A}_x v(x) &= \frac{nhf(x)}{\nu_0}\sum_{ij=11}^{rs}\sum_{i'j'=11}^{rs}\frac{nhf(x)}{\nu_0}\frac{(\hat{p}_{ij}(x)-\pi_{ij}(x))(\hat{p}_{i'j'}(x)-\pi_{i'j'}(x))}{\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)\pi_{i'\cdot}(x)\pi_{\cdot j'}(x)}}\\ &\quad\times\left(\delta_{i,i'}\delta_{j,j'} + 2\sqrt{\pi_{i\cdot}(x)\pi_{\cdot j}(x)\pi_{i'\cdot}(x)\pi_{\cdot j'}(x)} - \delta_{i,i'}\sqrt{\pi_{\cdot j}(x)\pi_{\cdot j'}(x)} - \delta_{j,j'}\sqrt{\pi_{i\cdot}(x)\pi_{i'\cdot}(x)}\right),\end{aligned}$$

which can be written, after some more algebraic work,

$$v(x)^t\mathcal{A}_x v(x) = \frac{nhf(x)}{\nu_0}\left(\sum_{ij=11}^{rs}\frac{(\hat{p}_{ij}(x)-\pi_{ij}(x))^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)} - \sum_{i=1}^{r}\frac{(\hat{p}_{i\cdot}(x)-\pi_{i\cdot}(x))^2}{\pi_{i\cdot}(x)} - \sum_{j=1}^{s}\frac{(\hat{p}_{\cdot j}(x)-\pi_{\cdot j}(x))^2}{\pi_{\cdot j}(x)}\right).$$

We can now write

$$\begin{aligned}\mathbb{V}\mathrm{ar}(v(x)^t \mathcal{A}_x v(x)) = \frac{(nh)^2 f^2(x)}{\nu_0^2} \Bigg(& \mathbb{V}\mathrm{ar}\left(\sum_{ij=11}^{rs} \frac{(\hat{p}_{ij}(x)-\pi_{ij}(x))^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}\right) \\ & + \mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{r} \frac{(\hat{p}_{i\cdot}(x)-\pi_{i\cdot}(x))^2}{\pi_{i\cdot}(x)}\right) + \mathbb{V}\mathrm{ar}\left(\sum_{j=1}^{s} \frac{(\hat{p}_{\cdot j}(x)-\pi_{\cdot j}(x))^2}{\pi_{\cdot j}(x)}\right) \\ & - 2\mathbb{C}\mathrm{ov}\left(\sum_{ij=11}^{rs} \frac{(\hat{p}_{ij}(x)-\pi_{ij}(x))^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}, \sum_{i=1}^{r} \frac{(\hat{p}_{i\cdot}(x)-\pi_{i\cdot}(x))^2}{\pi_{i\cdot}(x)}\right) \\ & - 2\mathbb{C}\mathrm{ov}\left(\sum_{ij=11}^{rs} \frac{(\hat{p}_{ij}(x)-\pi_{ij}(x))^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}, \sum_{j=1}^{s} \frac{(\hat{p}_{\cdot j}(x)-\pi_{\cdot j}(x))^2}{\pi_{\cdot j}(x)}\right) \\ & + 2\mathbb{C}\mathrm{ov}\left(\sum_{i=1}^{r} \frac{(\hat{p}_{i\cdot}(x)-\pi_{i\cdot}(x))^2}{\pi_{i\cdot}(x)}, \sum_{j=1}^{s} \frac{(\hat{p}_{\cdot j}(x)-\pi_{\cdot j}(x))^2}{\pi_{\cdot j}(x)}\right)\Bigg).\end{aligned}$$

The first variance term can be written as

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\sum_{ij=11}^{rs} \frac{(\hat{p}_{ij}(x)-\pi_{ij}(x))^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}\right) = & \sum_{ij=11}^{rs}\sum_{i'j'=11}^{rs} \frac{1}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)\pi_{i'\cdot}(x)\pi_{\cdot j'}(x)} \\ & \times \mathbb{C}\mathrm{ov}((\hat{p}_{ij}(x)-\pi_{ij}(x))^2, (\hat{p}_{i'j'}(x)-\pi_{i'j'}(x))^2),\end{aligned}$$

so that it can easily be derived from Corollary A.2(ii) that

$$\mathbb{V}\mathrm{ar}\left(\sum_{ij=11}^{rs} \frac{(\hat{p}_{ij}(x)-\pi_{ij}(x))^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}\right) = \frac{2\nu_0^2}{(nh)^2 f^2(x)}(rs-1) + O((nh)^{-3}) + O(n^{-1}h^3).$$

Similarly, one also finds

$$\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{r} \frac{(\hat{p}_{i\cdot}(x)-\pi_{i\cdot}(x))^2}{\pi_{i\cdot}(x)}\right) = \frac{2\nu_0^2}{(nh)^2 f^2(x)}(r-1) + O((nh)^{-3}) + O(n^{-1}h^3)$$

$$\mathbb{V}\mathrm{ar}\left(\sum_{j=1}^{s} \frac{(\hat{p}_{\cdot j}(x)-\pi_{\cdot j}(x))^2}{\pi_{\cdot j}(x)}\right) = \frac{2\nu_0^2}{(nh)^2 f^2(x)}(s-1) + O((nh)^{-3}) + O(n^{-1}h^3)$$

$$\mathbb{C}\mathrm{ov}\left(\sum_{ij=11}^{rs} \frac{(\hat{p}_{ij}(x)-\pi_{ij}(x))^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}, \sum_{i=1}^{r} \frac{(\hat{p}_{i\cdot}(x)-\pi_{i\cdot}(x))^2}{\pi_{i\cdot}(x)}\right) = \frac{2\nu_0^2}{(nh)^2 f^2(x)}(r-1) + O((nh)^{-3}) + O(n^{-1}h^3)$$

$$\mathbb{C}\mathrm{ov}\left(\sum_{ij=11}^{rs} \frac{(\hat{p}_{ij}(x)-\pi_{ij}(x))^2}{\pi_{i\cdot}(x)\pi_{\cdot j}(x)}, \sum_{j=1}^{s} \frac{(\hat{p}_{\cdot j}(x)-\pi_{\cdot j}(x))^2}{\pi_{\cdot j}(x)}\right) = \frac{2\nu_0^2}{(nh)^2 f^2(x)}(s-1) + O((nh)^{-3}) + O(n^{-1}h^3)$$

$$\mathbb{C}\mathrm{ov}\left(\sum_{i=1}^{r} \frac{(\hat{p}_{i\cdot}(x)-\pi_{i\cdot}(x))^2}{\pi_{i\cdot}(x)}, \sum_{j=1}^{s} \frac{(\hat{p}_{\cdot j}(x)-\pi_{\cdot j}(x))^2}{\pi_{\cdot j}(x)}\right) = O((nh)^{-3}) + O(n^{-1}h^3).$$

Then, it remains

$$\mathbb{V}\mathrm{ar}(v(x)^t \mathcal{A}_x v(x)) = 2(r-1)(s-1) + O((nh)^{-1}) + O(nh^5),$$

and from (A.15)

$$\mathbb{V}\mathrm{ar}(\hat{v}(x)^t \hat{v}(x)) = 2(r-1)(s-1) + O((nh)^{-1/2}) + O(nh^5).$$

Finally, as the order of any remainder term appearing in this development directly follows from the results of Lemma A.1 or Corollaries A.1–A.2, and that those results hold uniformly in $x \in S_X^h$, the same conclusion applies to these results, since all denominators appearing in the development is uniformly bounded away from zero with the considered assumptions. As the expression of the covariance is derived exactly the same way, in the same spirit as the proof of Lemma 2.1, it is omitted. □

Proof of Lemma 3.3

Remind that

$$V^2 = \frac{1}{n}\sum_k \|\hat{v}(X_k)\|^2 \mathbb{1}_{\{X_k \in S_X^h\}}.$$

Therefore, we have

$$\begin{aligned}\mathbb{E}(V^2) &= \mathbb{E}\left(\|\hat{v}(X)\|^2 \mathbb{1}_{\{X\in S_X^h\}}\right)\\ &= \mathbb{E}\left(\mathbb{E}\left(\|\hat{v}(X)\|^2 | X, X\in S_X^h\right)|X\in S_X^h\right)\mathbb{P}(X\in S_X^h).\end{aligned}$$

From Lemma 3.2, we have that

$$\mathbb{E}\left(\|\hat{v}(X)\|^2 | X, X\in S_X^h\right) = (r-1)(s-1) + O((nh)^{-1/2}) + O(nh^5)$$

where the remainder terms uniformly hold for any possible value of X in S_X^h, and as $\mathbb{P}(X\in S_X^h) = 1 - O(h)$, we conclude that

$$\begin{aligned}\mathbb{E}(V^2) &= (r-1)(s-1) + O((nh)^{-1/2}) + O(nh^5) + O(h)\\ &= (r-1)(s-1) + o(h^{1/2}),\end{aligned}$$

with the restrictions made on the bandwidth.

Now, write

$$\begin{aligned}\mathbb{V}\mathrm{ar}(V^2) &= \frac{1}{n^2}\sum_k\sum_{k'} \mathbb{C}\mathrm{ov}\left(\|\hat{v}(X_k)\|^2 \mathbb{1}_{\{X_k\in S_X^h\}}, \|\hat{v}(X_{k'})\|^2 \mathbb{1}_{\{X_{k'}\in S_X^h\}}\right)\\ &= \mathbb{C}\mathrm{ov}\left(\|\hat{v}(X_1)\|^2 \mathbb{1}_{\{X_1\in S_X^h\}}, \|\hat{v}(X_2)\|^2 \mathbb{1}_{\{X_2\in S_X^h\}}\right) + O(n^{-1}).\end{aligned}$$

This covariance can be written

$$\begin{aligned}&\mathbb{E}\left(\mathbb{C}\mathrm{ov}\left(\|\hat{v}(X_1)\|^2, \|\hat{v}(X_2)\|^2 | X_1, X_2, X_1\in S_X^h, X_2\in S_X^h\right)|X_1\in S_X^h, X_2\in S_X^h\right)\mathbb{P}(X_1\in S_X^h)^2\\ &\quad + \mathbb{C}\mathrm{ov}\left(\mathbb{E}\left(\|\hat{v}(X_1)\|^2 \mathbb{1}_{\{X_1\in S_X^h\}}|X_1\right), \mathbb{E}\left(\|\hat{v}(X_2)\|^2 \mathbb{1}_{\{X_2\in S_X^h\}}|X_2\right)\right).\end{aligned}$$

The second term is zero, as X_1 and X_2 are independent, and it follows, from (3.8), that

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(\|\hat{v}(X_1)\|^2, \|\hat{v}(X_2)\|^2\right) = \frac{2(r-1)(s-1)}{\nu_0^2}&\mathbb{E}\left(\nu_0^2\left(\frac{X_1-X_2}{h}\right)\middle| X_1\in S_X^h, X_2\in S_X^h\right)\\ &\times (1 + O((nh)^{-1/2}) + O(nh^5))(1-O(h)).\end{aligned}$$

See that $\mathbb{E}\left(\nu_0^2\left(\frac{X_1-X_2}{h}\right)|X_1\in S_X^h, X_2\in S_X^h\right) = \mathbb{E}\left(\nu_0^2\left(\frac{X_1-X_2}{h}\right)\right)(1+O(h))$, and that this expectation can be computed as

$$\begin{aligned}\mathbb{E}\left(\nu_0^2\left(\frac{X_1-X_2}{h}\right)\right) &= \iint \nu_0^2\left(\frac{x_1-x_2}{h}\right) f(x_1) f(x_2)\mathrm{d}x_1\mathrm{d}x_2\\ &= h\iint \nu_0^2(\delta) f(x_1 - h\delta) f(x_1)\mathrm{d}\delta\mathrm{d}x_1\\ &= h\int f^2(x_1)\mathrm{d}x_1 \int \nu_0^2(\delta)\mathrm{d}\delta + O(h^2)\\ &= hN_0\phi_0 + O(h^2)\end{aligned}$$

as $n\to\infty$. Finally, we find

$$\mathbb{V}\mathrm{ar}(V^2) = \frac{2(r-1)(s-1)}{\nu_0^2}(hN_0\phi_0 + O(h^2))(1 + O((nh)^{-1/2}) + O(nh^5))(1-O(h)) + O(n^{-1}), \tag{A.16}$$

that is the announced result. □

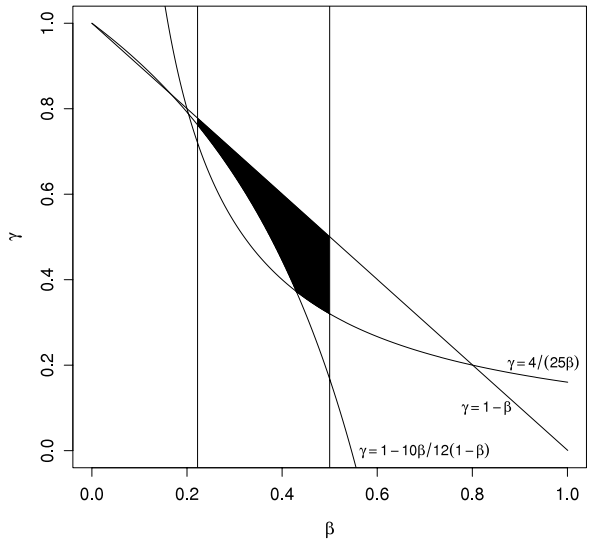

Fig. 2. Existence of parameter γ.

Proof of Theorem 3.1

Denote $\{\upsilon_{n,k}\} = \{\|\hat{v}(X_{(k)})\|^2 - \mathbb{E}(\|\hat{v}(X_{(k)})\|^2)\}$, where $\{X_{(k)}\}$ is the ordered version of $\{X_k\}$. Then, $\{\upsilon_{n,k}\}$ forms a triangular array of random variables such that for any n, we have a sequence of mean zero m_n-dependent random variables, with m_n growing to infinity. By Corollary A.1, we have that, for any k, $\mathbb{E}(\upsilon_{n,k}^{12})$ is bounded. Note $\Delta = \mathbb{E}(\upsilon_{n,k}^{12})$. Remember that we consider $h = O(n^{-\beta})$, for some $2/9 < \beta < 1/2$. Take some $\gamma \in [0, 1]$, such that $\gamma < 1 - \beta$, $\gamma > 1 - \frac{10\beta}{12(1-\beta)}$ and $\gamma \geq \frac{\alpha}{\beta}$, for some $\alpha \in]0, 4/25]$. Fig. 2 shows that there exists such γ, for any $\beta \in]2/9, 1/2[$. Now, Theorem 2.1 in [28] makes the job. Keeping our notation as close as possible as theirs, we hereafter verify their conditions (1)–(6), with m_n given by (3.9), and similarly to the derivation of (A.16),

$$B_{n,q_n,a}^2 \doteq \mathbb{V}\mathrm{ar}\left(\sum_{k=a}^{a+q_n-1} \upsilon_{n,k}\right) = 2hq_n^2(r-1)(s-1)\frac{\phi_0 N_0}{\nu_0^2} + o(hq_n^2) \quad \forall a$$

$$B_n^2 \doteq \mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} \upsilon_{n,k}\right) = 2hn^2(r-1)(s-1)\frac{\phi_0 N_0}{\nu_0^2} + o(hn^2).$$

Take $K_n = 2n^{1-\gamma}h(r-1)(s-1)\frac{\phi_0 N_0}{\nu_0^2} + o(n^{1-\gamma}h)$, $L_n = 2n^{1-\gamma-\alpha}h^{1-\gamma}(r-1)(s-1)\frac{\phi_0 N_0}{\nu_0^2}\frac{1}{4^\gamma \|f\|_\infty^\gamma} + o(n^{1-\gamma-\alpha}h^{1-\gamma})$ and see that, for n large enough,

$$\frac{B_{n,q_n,a}^2}{q_n^{1+\gamma}} = 2q_n^{1-\gamma}h(r-1)(s-1)\frac{\phi_0 N_0}{\nu_0^2} + o(q_n^{1-\gamma}h) \leq K_n \quad \forall a, \forall q_n \geq m_n$$

$$\frac{B_n^2}{nm_n^\gamma} = \frac{2(r-1)(s-1)\phi_0 N_0}{\nu_0^2 4^\gamma \|f\|_\infty^\gamma} n^{1-\gamma}h^{1-\gamma} + o(n^{1-\gamma}h^{1-\gamma}) \geq L_n.$$

Moreover, $K_n \to \infty$,

$$\frac{K_n}{L_n} = O(n^\alpha h^\gamma) = O(n^{\alpha-\beta\gamma}) = O(1)$$

as $\gamma \geq \frac{\alpha}{\beta}$,

$$\frac{\Delta}{L_n^6} = O(1),$$

as $\Delta < \infty$ and $L_n \to \infty$, and

$$\frac{m_n^{1+12(1+\gamma)/10}}{n} = O\left(n^{12(1-\gamma)/10}h^{1+12(1-\gamma)/10}\right) = o(1)$$

since $\gamma > 1 - \frac{10\beta}{12(1-\beta)}$. Thus, [28]'s CLT states that

$$\frac{\nu_0(1+o(1))}{n\sqrt{h}\sqrt{2(r-1)(s-1)\phi_0 N_0}} \sum_{k=1}^{n} \left(\|\hat{v}(X_k)\|^2 - \mathbb{E}\left(\|\hat{v}(X_k)\|^2 \right) \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

The announced result follows. □

References

[1] A. Agresti, An Introduction to Categorical Data Analysis, 2nd ed., in: Wiley Series in Probability and Statistics, 2007.
[2] B.S. Everitt, The Analysis of Contingency Tables, 2nd ed., Chapman and Hall, London, 1992.
[3] N. Mantel, W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease, J. Nat. Cancer Inst. 22 (1959) 719–748.
[4] P. McCullagh, J.A. Nelder, Generalized Linear Models, Chapman and Hall, London, 1989.
[5] G.F.V. Glonek, P. McCulagh, Multivariate logistic models, J. Roy. Statist. Soc. Ser. B 57 (1995) 533–546.
[6] G.F.V. Glonek, A class of regression models for multivariate categorical responses, Biometrika 83 (1996) 15–28.
[7] J.B. Lang, A. Agresti, Simultaneously modeling joint and marginal distributions of multivariate categorical responses, J. Amer. Statist. Assoc. 89 (1994) 625–632.
[8] J.B. Lang, J.W. McDonald, P.W.F. Smith, Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach, J. Amer. Statist. Assoc. 94 (1999) 1161–1171.
[9] K.Y. Liang, S.L. Zeger, B. Qaqish, Multivariate regression analyses for categorical data, J. Roy. Statist. Soc. Ser. B 54 (1992) 3–40.
[10] J.L. Horowitz, N.E. Savin, Binary response models: Logits, probits and semiparametrics, J. Econ. Persp. 15 (2001) 43–56.
[11] J.B. Copas, Plotting p against x, Appl. Stat. 32 (1983) 25–31.
[12] A. Azzalini, A.W. Bowman, W. Härdle, On the use of nonparametric regression for model checking, Biometrika 76 (1989) 1–11.
[13] M.C. Rodriguez-Campos, R. Cao-Abad, Nonparametric bootstrap confidence intervals for discrete regression functions, J. Econometrics 58 (1–2) (1993) 207–222.
[14] C.K. Chu, K.F. Cheng, Nonparametric regression estimates using misclassified binary responses, Biometrika 82 (1995) 315–325.
[15] D.F. Signorini, M.C. Jones, Kernel estimators for univariate binary regression, J. Amer. Statist. Assoc. 99 (2004) 119–126.
[16] J. Fan, I. Gijbels, Local Polynomial Modelling and its Applications, Chapman and Hall, 1996.
[17] G. Geenens, Non- and semiparametric models for conditional probabilities in two-way contingency tables, Ph.D. Diss., Institut de Statistique, Université catholique de Louvain, Belgium, 2008.
[18] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman and Hall, London, 1995.
[19] J. Rice, Boundary modification for boundary regression, Comm. Statist. A, Theory Methods 13 (1984) 893–900.
[20] R.L. Eubank, P.L. Speckman, Confidence bands in nonparametric regression, J. Amer. Statist. Assoc. 88 (1993) 1287–1301.
[21] Y. Xia, Bias-corrected confidence bands in nonparametric regression, J. Roy. Statist. Soc. Ser. B 60 (4) (1998) 797–811.
[22] M.C. Rodriguez-Campos, On confidence intervals in nonparametric binary regression via edgeworth expansions, J. Multivariate Anal. 69 (1999) 218–241.
[23] P. Hall, On bootstrap confidence intervals in nonparametric regression, Ann. Statist. 20 (2) (1992) 695–711.
[24] M.H. Neumann, Automatic bandwidth choice and confidence intervals in nonparametric regression, Ann. Statist. 23 (6) (1995) 1937–1959.
[25] D. Ruppert, S.J. Sheather, M.P. Wand, An effective bandwidth selector for local least squares regression, J. Amer. Statist. Assoc. 90 (1995) 1257–1270.
[26] S. le Cessie, J.C. van Houwelingen, A goodness-of-fit test for binary regression models, based on smoothing methods, Biometrics 47 (1991) 1267–1282.
[27] G. Geenens, Explicit formula for asymptotic higher moments of the Nadaraya–Watson estimator, 2008 (submitted for publication).
[28] J.P. Romano, M. Wolf, A more general central limit theorem for m-dependent random variables with unbounded m, Statist. Probab. Lett. 47 (2000) 115–124.